IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Appl. No.	:	10/823,829	Confirmation No. 5642
Applicant	:	Evans et al.	
Filed	:	April 14, 2004	
Art Unit	:	3714	
Examiner	:	Aileen Chyn	
Docket No.	:	CHORUS-P007-01	
Customer No.	:	27268	

SECOND DECLARATION OF ALAN L. COLQUITT

I, Alan L. Colquitt, Ph.D. of 7805 Timber Run Lane, Indianapolis, IN 46256 declare as follows:

1. I have over 20 years of experience implementing testing programs as an internal consultant in The Procter & Gamble Company (1985-1990) and Eli Lilly and Company (1990-present). I have a Ph.D. in Industrial and Organizational Psychology from Wayne State University in Detroit, MI. I have specialized training in the areas of test development, test validation, and psychometrics. I have considerable experience developing tests and testing programs for a variety of purposes including: Pre-employment screening; employee and leadership development; identification, selection and development of leaders and high potential employees; promotion; and reallocation. See attached resume for additional background (Exhibit A).

2. I have studied the disclosure of the above-identified patent application ("Evans") and the disclosures of U.S. Patent Publication No. 2002/0045154Q1 to Wood (The "Wood" application) and U.S. Patent Nos 7,148,969 and 6,341,267 to Bonstetter and Taub (the "Bonstetter Patent" and "Taub Patent" respectively).

3. I have studied the Office Action issued by the patent examiner on April 18, 2007 in application serial number 10/823,829; and have the following comments on the assertions in the Office Action in paragraph number 36 from the Office Action:

4. The examiner asserts that Wood provides multiple tests. Regarding this assertion, there are many different classes of tests and types of tests within classes. They are not interchangeable. Under Wood's system, any test will do (e.g. Wood, p5 [0110]). The choice of tests needs to be specific to the purpose for which they are being used and not all tests relevant

for a given purpose are created with the same rigor and level of professional quality as standardized measures of psychological constructs. The claimed invention relates to personality attributes because personality may be modified and improved, whereas Wood relates to temperament attributes that are characteristics of individuals that are recognized by one of skill in this field as being innate, relatively unchangeable aspects of the individual. Wood refers frequently to 2 tests used in his system: The Keirsey Temperament Sorter and the Myers-Briggs Type Indicator. Neither of these tests is well-suited to providing feedback and development advice for rating competencies or creating transformational activities. First, these 2 instruments measure temperament, not personality. Temperament is felt to be permanent, not changeable (see Strelau, 1987, as cited in Hofstee, 1991, p182). They talk about type as like "handedness"; one is born with it. It makes little sense to use instruments that measure things that cannot be changed in a context where a method is employed to enable individuals to grow and develop competencies—to change. The tests recited in the claims measure Personality, which is felt to be more environmentally controlled and therefore can be developed. Second, the Wood "tests" were not developed to be used for human capital management. The Wood "tests" are not useful in a context where results need to be linked to competencies, where feedback is provided on how strong/weak someone is related to personality attributes or competencies. The language used in the Wood "tests" is not immediately relevant to assessing competencies. Respondents are classified as "ENTJ"—one of the MBTI types, or "Guardians", "Artisans"—output from the Keirsey). It is not clear what these mean and how they fit in the context of assessing competencies without additional invention to help bridge this gap. The tests recited in the present claims were developed for assessing personality attributes and do not suffer from this issue. Third, the Wood "tests" identify a respondent's *type*. *Types* are not "good" or "bad". They do not indicate the strength of ability or personality or degree of aptitude. Because of this it is difficult to link test results to the strength or weakness of competencies. The standardized measures of the capability tests in the claims and disclosed in the specification have traditional scaling (scale scores produced with high and low scores relative to an external norm group) which lends itself much more to being linked to competencies. Finally, there are also technical problems with the way one of the Wood "tests" is scored and with its reliability and validity. These issues the "test" less useful psychometrically (see Hunsley, Lee, & Wood, 2004 for a discussion of issues with the Myers-Briggs Type Indicator).

5. The examiner asserts that "answering questions about finance is the same as taking a finance test and that discussions of "test development" are not relevant to this discussion. Regarding this first assertion, answering questions is not taking a test. Finding out

what someone knows about finance does not mean someone is taking a finance test. A finance test could involve someone verbally asking questions. However, a number of other criteria would need to be satisfied for this to be considered a test. As stated in Guion, 1998, p. 485:

> *"A test is an objective and standardized procedure for measuring a psychological construct using a sample of behavior. A test is objective in that responses can be evaluated against external standards of truth or of quality, correct or incorrect, or better or poorer than a standard. Measurement implies quantification. Tests are scored quantitatively, with measurable precision, on numerical scales representing levels of a construct to be inferred from the scores. Tests use a standardized procedure with the same stimulus component for all test takers. Standardization refers primarily to controlling the conditions and procedures of test administration."*

One of ordinary skill in the art would not consider to be a test what Wood refers to as "asking questions about finance". This does not meet the above conditions and standards. Regarding Wood's use of the word "test" in the abstract, Wood is free to refer to the tools used in his system by any term he chooses. One of ordinary skill in the art would use more precision than is used by Wood in distinguishing the different types of assessment procedures. For some of the "tests" he describes, those skilled in the art would properly refer to them as tests or inventories, and they would meet the conditions describe in Guion, 1998. The process he describes with respect to "asking questions about finance" would not qualify using the standards described in Guion, 1998. Those skilled in the art would not refer to this as a test.

6. The examiner asserts that the techniques described in Wood are similar to those in the claimed invention and that the techniques provided in Bonstetter would have been obvious additions to the system described by Wood. Regarding this assertion, Bonstetter is solving a different problem than the claimed invention. The problem Bonstetter identifies is the high failure-rate in the ability to identify job candidates who will be successful on the job (Column 1, Paragraph 2, lines 25-35). His solution to this problem is to improve the way competencies are IDENTIFIED (not measured) (see column 3, prior to Section III.) His solution is not to measure people on competencies. The problem the claimed invention addresses is how to measure people on competencies using commercially available and professionally developed and validated tests, combining the information together to help individuals improve. Bonstetter comments that the reasons selection interviews are not effective is that interviewers don't focus on the right things. His process IDENTIFIES those "right things"-- the job-related requirements for a job. Those skilled in the art refer to this process as "Job Analysis". (See Harvey, 1991, p. 74). The output

of a job analysis serves as input to a whole host of processes, one of which might be the development of selection procedures such as the selection interview as described in Bonstetter (see Harvey, 1991, pp. 124-146 for the varied uses of job analysis information). Bonstetter is focused on the job and its requirements. The claimed invention tests the individual and links those measurements to competencies to help the individuals develop these competencies. The examiner asserts that the constructive feedback provided by Bonstetter is the same as the feedback provided by the claimed invention. Regarding this assertion, there are several issues with what Bonstetter describes and there are several ways it differs from the claimed invention. As an example, Bonstetter shows a Personal Competency Inventory (PCI) in Figure 12A. He reports that participants and others complete this form, rating competencies directly. The output is a set of strengths (see Fig.14b). First, it is not clear where Bonstetter's system ends and where other secondary uses of his system begin. The examiner refers to one of the possible uses of Bonstetter's system and the information the system produces. The examiner's statement "match people to appropriate jobs" refers to using the information for selection interviewing. The examiner's statement "improve people's performance on the job by giving them feedback on their strengths and weaknesses which pertain to job performance" refers to using the information to develop a performance appraisal system or a training and development/feedback and coaching system. The Bonstetter system focuses on the position. He says in Col 5, 35-40 (referring to Figure 12a) this is something that CAN be used with the invention, and in Col 5, 44 that Figures 14A and 14b are HYPOTHETICAL examples of reports. Points 9, 13, 15, and 16-35 under Columns 3 and 4 all refer to other uses for the information his system produces. His system does not include the procedures to accomplish these "other uses." He says his system "allows for" other things to be done (e.g. "This allows a set of interview questions to be produced"), but these things are not part of his system. The primary purpose for which he uses the information from his system is to "evaluate applicants for the position to determine if their characteristics will make them high performers in the position" (Col 4, paragraph 30). In contrast to Bonstetter's job analysis, the claimed invention measures individuals on competencies and gives them feedback and ideas for development. Second, Bonstetter presents a tool in Figure 12a which could be used to measure competencies. This is something he himself laments may be difficult to do based on what is currently known about competencies:

> *"The many attempts....beg the question—how does one define 'competencies' and which ones are relevant? There is no agreement on these questions."* (Col 2, lines 43-45). Additionally he reports "*As previously discussed, much has been written about 'competencies'. However, no agreement exists as to what is a competency* (Col 8, lines 48-51).

Those of ordinary skill in the art would agree it is difficult to measure competencies directly given their breadth. There are several issues with the tool shown in Figure 12a as it is shown. The tool does not appear to be a published tool, nor does it appear to be professionally developed and validated to provide a standardized measure. It purportedly measures competencies; however, it contains questions similar to those typically included in personality tests. As described in Spencer, McClelland and Spencer (1994) (cited in my original affidavit), personality is only one component of competencies. He provides no evidence this tool measures specific personality traits, nor does he list these traits or how these traits are linked to specific competencies. Moreover, there are far too few items in the test to reliably measure standard personality traits or to reliably measure the number of competencies listed. Third, while Bonstetter says his system provides feedback on competencies (assuming Figures 14a and 14b are part of his system), the form of this feedback is simply a rank ordering of competencies from highest to lowest. Candidates are not provided any detail about what specific attributes might be contributing to their strengths and weaknesses, nor do they have any information on how they might improve in these areas. The claimed invention provides a comprehensive individual capability evaluation that rates competencies, as well as providing information on how to improve in these areas. Finally, In Col 14, section F, Bonstetter discusses alternative features and options that might be obvious to one skilled in the art. These alternatives are no more obvious to those skilled in the art than are the many uses to which any job analysis information can be put (as articulated in Harvey, 1991, p125-146). Anyone of ordinary skill in the art who is familiar with job analysis knows the variety of ways the information it produces can be used. However, he/she may not possess the knowledge, skills, and experience to know how to develop or improve any of these systems. Specific innovations in the areas of assessment and training and development or the development of a compensation system would not be obvious to those conducting a job analysis. For example, the fact that someone conducts a job analysis of a sales position does not mean it would be obvious to that person how to develop or improve a sales incentive compensation system (which would use the job analysis information as input).

7. The examiner asserts that Wood measures competencies and that personality traits *are* competencies. Regarding this assertion, first, the psychological or temperament constructs measured in the tests used by Wood are not competencies. Introversion/Extraversion is not a competency (see Spencer, McClelland, and Spencer, 1994 in the original affidavit for a definition of competency). Measuring someone's type (e.g. ESTJ or Artisan) is not assessing a competency. In fact, Wood himself distinguishes instruments that measure competencies from instruments that measure other psychological constructs (see p.5, [0109] to [0125]. The

definitions cited in Ash et al, 2000 from my prior affidavit suggest that a competency is not a unitary concept. Competencies can be, in part, defined by personality traits, but would also be defined by other constructs such as skills and abilities (See Spencer, McClelland, and Spencer (1994) in original affidavit). They are not substitutable for one another. Evidence about extraversion-introversion (a personality trait) may be used to assess whether or not someone has strong interpersonal competence....however, this does not mean introversion-extraversion is a competency. Second, the psychological constructs measured by the tests used in Wood are not personality attributes linked to competencies.

8. The examiner asserts that the systems of Wood and Bonstetter can be used to solve the same problems. Regarding this assertion, the problem Wood is addressing is how to segment people based on psychological instruments to better deliver products and services to customers. The information he uses to segment people are psychological tests. Wood outlines countless different services that could be provided, one of which is to help match people to jobs. However, these services are not an essential part of Wood's invention. Wood's system focused on the psychological tests and the segmentation. The problem Bonstetter is addressing is how to obtain better information about what jobs require so he can better identify candidates who can be successful at these jobs. His primary focus is on profiling jobs and he acknowledges this information can be used in many different ways. Wood is focused on people and how to segment them; Bonstetter is focused on jobs and how to profile them. One of ordinary skill in the art would consider these to be two very different issues, neither of which is particularly useful for evaluating a plurality of competencies using a plurality of tests assessing attributes. As is detailed in (Guion, 1998, p. 57) the first step in the development of HR systems and processes is to understand the job (job analysis). This is where Bonstetter focuses. The second step is to translate the job requirements into people requirements. This is where Wood focuses. Neither Wood nor Bonstetter speak of rating a plurality of competencies or developing transformational activities.

And further, I sayeth not.

Date: July 17, 2008

Alan L. Colquitt

References

Exhibit A-- Resume of Alan L. Colquitt

Exhibit B--

Harvey, Robert J (1991). Job Analysis. In Dunnette, M. D. & Hough L. M. (1991) Handbook of Industrial and Organizational Psychology, Volume II. Consulting Psychologists Press, Inc.

Exhibit C--

Hofstee, W.K.B (1991). Personality and Temperament. In Strelau, J. & Angleitner, A. (1991). Explorations in Temperament: International Perspectives on Theory and Measurement. Plenum Press: New York

Exhibit D--

Hunsley et al., 2003 J. Hunsley, C.M. Lee and J.M. Wood, Controversial and questionable assessment techniques In: S.O. Lilienfeld, S.J. Lynn and J.M. Lohr, Editors, *Science and pseudoscience in clinical psychology*, Guilford Press, New York (2003), pp. 39–76.

References:

Alan L. Colquitt, Ph.D.
7805 Timber Run Lane
Indianapolis, Indiana 46256
acolquitt@comcast.net
(317) 849-1258

EDUCATION AND TRAINING

Doctor of Philosophy (December, 1986)
Wayne State University
Detroit, Michigan
Industrial/Organizational Psychology

Licensed Psychologist (1994 to present)
State of Indiana. License #20040761

Bachelor of Arts (May, 1982)
Indiana University
Bloomington, Indiana
Psychology

Advanced Organizational Development/Human Resources Development (OD/HRD) program (Spring, 1995). Columbia University,

PROFESSIONAL EXPERIENCE

Eli Lilly and Company
Workforce Research
Manager (January 1997 to present)

Key Areas of responsibility:

- Develop, implement a variety of survey processes focused on key stakeholders (employees, collaborators)
- Develop, implement, and monitor testing and assessment systems
- Conduct special research on issues of current interest (e.g. culture, diversity issues, retention, impact of key interventions, diagnose specific issues)
- Develop, implement, and report key workforce and people-related metrics
- Leverage results and learnings from all research, monitor trends affecting workforce
- Consult with line, HR management in the above subject areas
- Manage small staff (1 person) of technical experts

Eli Lilly and Company
Organization Effectiveness
Manager (June 1994 to January 1997

Key areas of responsibility:

- Organizational consulting
- Organization design
- Organization change management
- Strategy development

- Team development
- Workforce research, strategic studies
- Survey development and design
- Performance management system design
- Testing and assessment program development

Eli Lilly and Company
Career Development and Psychological Services
Staff Psychologist (September 1990 to June 1994

Key areas of responsibility:

- Assessment and psychometrics
- Design and administer career assessment and development programs
- Testing and assessment program development
- Workforce research, strategic studies

The Procter & Gamble Company
Personnel Research
Manager (December 1985 to September 1990)

Key areas of responsibility:

- Testing and assessment program development
- Test development and validation
- Training evaluation
- Strategic studies and research

AREAS OF TECHNICAL COMPETENCE/EXPERTISE

- Assessment and selection systems (tests, simulations, assessment centers, test validation, psychometrics)
- Survey design, development
- Program evaluation
- Performance management/Performance appraisal
- Job analysis, needs analysis
- Research design
- Statistical analysis
- Career and employee development
- Performance measurement, metrics
- Organization diagnosis
- Organization change management
- Organization design
- Strategy development

PUBLICATIONS AND PRESENTATIONS

Colquitt, A.L. (2008). Total Rewards at Eli Lilly and Company: Applying Total Rewards Optimization. Paper presented as a part of symposium: "Optimizing HR: Tracking the return on investments in people." Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L., Fink A., Futrell D.A., and Johnson S. (2008). More survey ponderables…Questions and Answers on Effective Employee Surveys. Annual meeting of the IO/OB conference, Indianapolis, IN

Colquitt, A.L., & Futrell D. A. (2007). Questions and answers about survey research: Lessons learned from survey programs at Eli Lilly and Company. Annual meeting of the IO/OB conference, Indianapolis, IN

Colquitt, A.L., Mastrangelo, P., and Weiner, S. (2006). Staying on your high horse: Ethical challenges in employee surveys. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. & Macey W. H. (2005). Surveys throughout the employment lifecycle: What matters when. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. & Futrell, D.F. (2004). Use of a biodata selection instrument to improve retention. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. & Lange C. (2004). Gender diversity at Eli Lilly and Company: Follow-up on the "Leaders in a Global Economy" study. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. (2003). Working inside on the balanced scorecard: Lessons learned about strategy, tactics, and culture. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. (2002). Getting systematic about retention in one company: Strategy, Tactics, and Learnings. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. (2001). After the Rating Stops: Effecting Change with Multi-Source Feedback. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. (2000). Predictors of Turnover for Sales Representatives: The "Fruits" of an Exit Survey Process. Annual meeting of the Society of Industrial and Organizational Psychologists

Colquitt, A.L. & Futrell D.A. (2000). Automated Technologies for Biodata Prediction Systems. Annual meeting of the Society of Industrial and Organizational Psychologists

Becker, T & Colquitt A. (1992). Potential vs. Actual faking of a biodata form: An analysis along several dimensions of item types. Personnel Psychology, 45, 389-406

PROFESSIONAL AFFILIATIONS

Society for Industrial and Organizational Psychology
The Mayflower Group (survey consortium)
American Psychological Association
Academy of Management
Human Resource Planning Society

HONORS AND DISTINCTIONS

1997-2000	The Mayflower Group board of governors, Chair in 2000	1978	Hoosier Scholar
1982	Phi Beta Kappa	1978	National Merit Scholar

REFERENCES

Available upon request

general situational constraints, including both lack of resources and lack of information, (b) supervisor support, (c) training or opportunity to use skills, (d) job or task importance, and (e) unit cohesion and peer support.

A General Approach to Need Analysis

Organizational need analysis is a managerial, not a research, function. Its immediate purpose is to generate hypotheses, not to test them. It must be done systematically, recognizing that the outcome of need analysis is a judgment (or a set of judgments) that can be framed in the language of hypotheses, and the quality of the judgment depends on the experience, knowledge, and wisdom of those who reach it. The best advice I can offer, whether the focus is on dialogue or on questionnaires, is to consider five general questions as carefully and with as much collaboration as possible:

1. What work outcomes are most in need of improvement? That is, what outcomes are most highly valued and not satisfactorily attained, and what ones are most deplored and too much in evidence? In either case, remedial action is needed; what priorities can be set for the importance of such remedies? Answers provide the criterion concepts for *any* hypotheses geared to improving the situation.
2. How widespread is the problem? Is it pervasive throughout an organization or organizational unit, or is it found in specific instances (i.e., specific people or specific units)?
3. At what level of analysis (organizational unit or individual) is the problem most accurately defined and approached? Consider, for example, a serious turnover problem. Should it be approached at the work unit level or the individual level or a broader organizational level?
4. What kinds of corrective actions are plausible? That is, what might reasonably, feasibly, be expected to help? What is the range or scope of sensible possibilities? Identifying a full range seems to call for the collective experience of people with a variety of backgrounds. Discussions with different people in the organization, and perhaps with outside consultants, can provide an initial list of plausible actions.
5. How effective have the various options been in prior use, in this organization or elsewhere? It is probably this question that gives some edge to attempts to improve individual selection decisions when the problem is one of improving performance levels; most other activities lack the strong research base, with the relatively substantial levels of predictive power and utility, that characterizes the testing literature.

Insiders—people who know the organization intimately—are necessary participants in seeking the answers; this is not something that a

manager can delegate to an outside consultant and merely await the report—although an outsider can facilitate the discovery of answers and the reduction of internal barriers to their expression. Collectively, participants must have a wide range of knowledge, of interest, and of technical expertise—more than is likely to be found in any one person. The best procedure for organizational need analysis may be to form a task force of bright people who know the organization from a variety of perspectives, augment them as necessary with hired specialists in various problem solutions or in discussion processes, and let them study, question, argue, and arrive at their best collective judgments.

JOB ANALYSIS

When organizational needs require improved personnel decisions for people on specific positions, jobs, or groups of jobs, job analysis (or position analysis) is necessary. Jobs are analyzed to understand them clearly enough to know which variables or performance constructs should be predicted and to identify variables or constructs that might be effective predictors—that is, to develop predictive hypotheses.

Some Definitions

Following McCormick (1979), with some additions and liberties of my own, here are some more or less standard definitions:

Position: The duties and tasks carried out by one person. A position may exist even where no incumbent fills it; it may be an open position. There are at least as many positions in an organization as there are people.

Job: A group of positions with the same major duties or tasks; if the positions are not identical, the similarity is great enough to justify grouping them. A job is a set of tasks within a single organization or organizational unit. This definition does not preclude flexibility. Members of a self-contained work unit may, on any given day, be doing different tasks, but each member may also be expected to do on another day any task the group as a whole must do.

Occupation: An occupation is a class of roughly similar jobs, found in many organizations and even in different industries. Examples include attorney, computer programmer, mechanic, and gardener.

Job family: A group of jobs similar in specifiable ways, such as patterns of purposes, behaviors, or worker attributes. Pearlman (1980) applied

Exhibit B

the *family* concept to occupations, but the term is usually applied to sets of jobs within an organization.

Job analysis: Job analysis is a study of what a jobholder does on the job, what must be known in order to do it, what resources are used in doing it, and perhaps the conditions under which it is done. What the jobholder does may be defined in several ways: as tasks, classes of duties or responsibilities, broad activities, or general patterns of behavior. What must be known includes job knowledge and job skills. What resources are used may include those the person may bring to the job (relevant experiences, general abilities, or other personal characteristics), tools and materials used (e.g., manuals or handbooks, supplies, or equipment) or the work products of other jobs or work units.

Element: The smallest feasible part of an activity or broader category of behavior or work done. It might be an elemental motion, a part of a task, or a broader behavioral category; there is little consistency in meanings of this term.

Task: A step or component in the performance of a duty or activity. A task has a clear beginning and ending; it can usually be described with a brief statement consisting of an action verb and a further phrase.

Activity (or *responsibility* or *duty*): A relatively large part of the work done in a position or job. It consists of several tasks related in time, sequence, outcome, or objective. A clerical example might be "sorting correspondence" or "handling cash" or "preparing reports." All tasks grouped under these activities are done for a common end. One task in correspondence sorting might be "identify letters requiring immediate response." Putting together a report includes such tasks as laying out or formatting tables and charts, typing text, typing tables or charts, proofing for errors, and perhaps duplicating, collating, and binding copies of the reports. Activities and tasks are both components of jobs, but activities are usually considered more general, more encompassing.

Essential function: A term introduced in the *Americans With Disabilities Act (ADA)*, which defines a "qualified individual with a disability" in part as one who "can perform the essential functions of the employment position that such an individual holds or desires" (Schneid, 1992, p. 28). The meaning of many terms in the ADA, including this one, waits on court decisions and developing case law. In the meantime, EEOC regulations identify three considerations: (a) whether the position exists for the purpose of carrying out the function, (b) whether the number of employees who can perform the function is limited, and (c) whether the function is highly specialized so that people are

hired because of their special expertise or ability to carry out the function (Schneid, 1992, pp. 33–34).

Job description: A written report of the results of job analysis. A job description is usually narrative, sometimes given in a brief summarizing paragraph. It may be more detailed. Where job analysis was done by survey methods, the description may include listings of task statements found to define or characterize the job being studied, along with statistical data.[3]

Job specification: Required qualifications for the job (or position), as revealed in the job description. Depending on the job or job category, specifications can include legal requirements (age, licenses, residency, etc.), education, skills, or perhaps assessment standards (although the latter requires research beyond the job description).

Detail Versus Generality in Job Analysis

In job analysis, a job as a whole is analyzed into component parts; the level of detail can vary widely. Detailed statements may be best for developing training programs, but more general statements are more useful for identifying criteria and predictors for selection (Lawshe, 1987).

Clarity counts more than detail. Lawyers and courts want more detail than is useful. Too much detail can muddle matters; what is needed is a clear enough understanding of the job to move on to the next step, the development of one or more predictive hypotheses. These require the wisdom, insight, and even introspection of people who know and understand the job. Job analysis can tap the wisdom and knowledge of job experts. Highly detailed, cover-all-bases, formal job analysis may not be needed at all—except possibly for convincing others that the analysis was done well.

Information *needed* is not necessarily the information *desired*. In an age of litigation, actions are governed as much by what is prudently filed away as by what is actually needed. Fine details may not be needed for any purpose beyond a trial. Failure to convince a trial judge that the job analysis was "adequate"—lots of questions asked, results recorded in a lengthy job description along with lots of statistical analyses—may be the

[3] In a peculiar pair of definitions, the *Uniform Guidelines* defines job analysis as "a *detailed* statement of work behaviors. . . ." and job description as "a *general* statement of job duties. . . ." (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978, p. 38307, italics added). Obviously, these definitions were written without much regard for the meanings of the verbs *to analyze* or *to describe*. In my judgment, the failure to recognize a difference between the process of analyzing something and the description of the results has resulted in mischief in court cases.

Guion, 1998

11

Assessment by Testing

Of all assessment methods, testing has the best foundation in research, measurement theory, and the development of standards of evaluation. Other methods of assessment may be preferred for some purposes, but the development and use of tests provides a prototype for the development, use, and evaluation of assessment in other forms.

A *test* is an objective and standardized procedure for measuring a psychological construct using a sample of behavior. A test is objective in that responses can be evaluated against external standards of truth or of quality—correct or incorrect, or better or poorer than a standard. Measuring implies quantification. Tests are scored quantitatively, with measurable precision, on numerical scales representing levels of a construct to be inferred from the scores. A *construct*, as I use the term, is a fairly well-developed idea of a trait; most constructs in testing are abilities, skills, or areas of knowledge. Tests use a standardized procedure with the same stimulus component for all test takers. *Standardization* refers primarily to controlling the conditions and procedures of test administration, that is, keeping them constant—unvarying. If scores from different people are to be comparable, they must be obtained under comparable circumstances. If people tested in one room have 30 minutes in which to complete a test, and those in another have only 20 minutes, neither the circumstances nor the scores are comparable. Any circumstances of test administration potentially influencing scores should be standardized. More than anything else, it is attention to standard procedure that distinguishes testing from other forms of assessment. The distinction is fuzzy. In this chapter, I describe a variety of procedures for assessing KSAs,

ranging from highly standardized tests to assessments with little or no standardization, with no clear line distinguishing tests from other assessments procedures.

Defining a test as a sample of behavior means that the examinee is not passive but does something. In other kinds of testing (e.g., blood tests) the object of measurement sits passively while something is done to it. In psychological tests, the examinee responds to test stimuli by writing answers to questions, choosing among options, recognizing or matching stimuli, performing tasks, ordering objects or ideas, or producing ideas to fit requirements—and this is not an exhaustive list.

TRADITIONAL COGNITIVE TESTS

Cognitive tests allow a person to show what he or she knows, perceives, remembers, understands, or can work with mentally. They include problem identification, problem-solving tasks, perceptual (not sensory) skills, the development or evaluation of ideas, and remembering what one has learned through general experience or specific training. They include intelligence tests, achievement tests, and job knowledge tests, among others.

Some History: Oral Trade Tests

Oral trade tests were among the earliest of employment tests. Many were developed in the Army during World War I and in industrial locations in the years immediately after (Chapman, 1921; Link, 1919; Poffenberger, 1927). Trade tests were needed, in part because many applicants genuinely thought they had expertise in a trade when, in fact, they knew only a limited aspect of it; oral tests were needed because many applicants could not read. The problems increased in the depression years and were addressed with excellent test development work in the United States Employment Service (USES). The work, described by Osborne (1940), is unfortunately no longer well-known.

USES job analysts in different regions made detailed observations and developed questions about a wide range of topics fundamental to the trade. Questions and the correct answers were reviewed regionally and nationally for adherence to specified principles, such as brevity. Research compared scores in three groups of subjects. People in the *A* group were highly skilled in the trade with at least four years of post-training experience. The *B* group included apprentices, helpers, or beginners. The *C* group consisted of people in other occupations whose work gave them contact with workers in the targeted trade. The proportion of correct

FORM I

Score	Expert Bricklayers (n = 65)	Apprentices and Helpers (n = 25)	Related Workers (n = 35)
15	xx		
14	xxxxxxxx		
13	xxxxxxxxxxxxxxxx		
12	xxxxxxxxxxxxxxxxxxxx *		
11	xxxxxxxx	x	
10	xxxxx	xx	
9	xx		x
8	xx	x	
7		xx	
6		x	x
5		xxxxx *	
4		xxxxx	x
3		xx	xxx
2		xx	xx
1		xxx	xxxxxxxxxx *
0			xxxxxxxxxxxxx

FORM II

Score	Expert Bricklayers (n = 65)	Apprentices and Helpers (n = 25)	Related Workers (n = 35)
15	x		
14	xxx		
13	xxxxxxx		
12	xxxxxxx		
11	xxxxxxxxxx		
10	xxxxxxxxxxxx *		
9	xxxxxxxxxxx		
8	xxxxx	x	x
7	xxxx		x
6	x	x	x
5		xx	
4	x	xxx	x
3		xxxxx	xxx
2	x	xxxxx *	xxxxx
1		xxxx	xxxxxxxxxx *
0		xxxx	xxxxxxxxxxx

* median score

FIG. 11.1. Distribution of scores on two forms of the oral trade test for bricklayers. From Osborne (1940).

answers should be significantly higher in group *A* than in groups B or C.[1] An example of the results for bricklayers is shown in Fig. 11.1.

Traditional Tests

Most tests now used are called paper-and-pencil tests, but materials do not define traditional tests. The defining features of traditional tests are that they are well-standardized, that their items can be reliably scored, and that they can be administered to groups of people.

[1]Later, as experience demonstrated that response patterns varied little between groups *B* and C, only one "control group" was used. It was designated group C, mainly apprentices, helpers, or beginners but augmented as needed by workers in related occupations.

Dunnette & Hough (1991)

context. However, much of our current semantic confusion is self-inflicted, being caused by inconsistent usage of terms in the research literature, texts, and relevant professional standards documents. Surprisingly, one of the most troublesome issues concerns the definition of job analysis.

What is Job Analysis? The McCormick (1976) chapter is perhaps the most widely cited and influential recent treatment of this topic. McCormick (pp. 652–653) defined job analysis as the collection of data on (a) "job-oriented" behaviors, such as job tasks and work procedures; (b) more abstract "worker-oriented" behaviors, such as decision making, supervision, and information processing; (c) behaviors involved in interactions with machines, materials, and tools; (d) methods of evaluating performance, such as productivity and error rates; (e) job context, such as working conditions and type of compensation system; and (f) personnel requirements, such as skills, physical abilities, and personality traits.

Building on McCormick's definition, I define *job analysis* as the collection of data describing (a) observable (or otherwise verifiable) job behaviors performed by workers, including both *what is accomplished* as well as *what technologies are employed* to accomplish the end results and (b) verifiable characteristics of the job environment with which workers interact, including physical, mechanical, social, and informational elements. Job behaviors or contextual characteristics can be observed both directly (e.g., physical actions performed, tools and machines used, people contacted, materials modified, services provided, or sources of data used as input) as well as indirectly, through the use of what McCormick et al. (1972, p. 384) termed *strong inference* from other observable job behaviors. For example, an activity like making managerial decisions regarding investments and cash flow may be described in terms of the work products involved or produced by the decision, the kinds of data used or produced by the decision maker, or the consequences of incorrect decisions.

My definition of job analysis parallels McCormick's by emphasizing the description of work behaviors, work products, and job context, including the informational and social context in which the work is performed. It differs, however, by excluding the process of inferring hypothetical worker traits or abilities (variously termed *job specifications*, *worker specifications*, or *personnel specifications*) presumed to be required for job performance (e.g., "general cognitive ability," "leadership," "dominance," "sociability"). Inferring required personal traits—though admittedly an important part of the employee selection process—is excluded because it fails to meet the following three criteria that I maintain should characterize a job analysis method.

First, job analysis methods should have as their goal the description of *observables*. This view parallels the position taken in the *Uniform Guidelines on Employee Selection Procedures* (1978, often termed the *Guidelines* or *Uniform Guidelines*), particularly regarding the meaning of the term *work behavior*:

> Job analysis ... includes an analysis of the important work behavior(s) required for successful performance and their relative importance and, if the behavior results in work product(s), an analysis of the work product(s). Any job analysis should focus on the work behavior(s) and the tasks associated with them. If work behavior(s) are not observable, the job analysis should identify and analyze those aspects of the behavior(s) that can be observed and the observed work products. (section 1607.14.C.2)

Thus, even behaviors that some might consider unobservable, such as coordinating, deciding, or planning, can be objectively described in terms of the observables that result from their performance, such as people contacted in the course of making the decision, kinds of

[illegible] ordered, or consequences of the decision. All work behaviors, no matter how abstract they may appear at first, must have an observable component and will eventually result in an observable action, state, or product.

If a work behavior does *not* have an observable component, then by definition it cannot be described in a job analysis. As the *Guidelines* note, an *observable* is something able to be seen, heard, or otherwise perceived by a person other than the person performing the action" (section 1607.16.N). Describing observables should be the sole goal of a job analysis.

Second, a job analysis should involve the description of work behavior *independent of the personal characteristics or attributes of the employees who perform the job*. That is, the work itself is being described, not the personal traits or performance effectiveness of the people who currently attempt to perform the job. A job analysis describes *how* a job is performed and must not be colored by whether or not the employees currently hired to perform the job are doing so successfully or not. The unit of analysis in job analysis is the *job* or the *position* (defined below), *not* the *incumbents* who perform the work. Techniques that describe the level of effectiveness of individual employees are called *performance appraisals*, not job analyses.

Third, and of critical importance, job analysis data must be *verifiable* and *replicable*. That is, should the accuracy or validity of a job analysis be challenged, the organization must be able to justify every job analysis rating in terms of observable behaviors, actions, work products, information used or produced, and the like. The importance of restricting job analysis to the domain of verifiable observables cannot be overemphasized. Key to the ability to demonstrate the validity of job analysis data is the requirement of *replicability*—that is, independent observers with adequate job familiarity should produce functionally equivalent ratings.

Obviously, some degree of inference is involved in most rating processes; to the extent they are present, inferences in job analysis should be of the "strong" variety defined by McCormick et al. (1972). That is, less observable behaviors (e.g., "making decisions") can be described in terms of their observable, manifest aspects. "Weak" inferences—in particular, inferring the hypothetical human constructs presumed necessary for personal success on the job—should not be termed job analysis.

Knowledge/Skills Versus Abilities/"Others" and Job Specifications Versus Job Analyses. Other than the term job analysis, the phrase *job knowledge, skills, abilities, and other characteristics* (KSAs or KSAOs) has probably caused more semantic confusion than any other. Because of the fundamental differences that exist between job knowledge and skill (KS) versus ability and "other" (AO) specifications, I find it much more useful to replace the term KSA/KSAO with a discussion of KS versus AO requirements.

When the component K, S, A, and O terms are defined, it is readily apparent that KSs are specified directly in terms of observable job behaviors, whereas AOs are only indirectly—if at all—linked to the actual job behaviors identified in a job analysis. In the case of KS requirements, the 1978 *Guidelines* define job *knowledge* as "a body of information applied directly to the performance of a function" (section 1607.16.M), and a job *skill* as "a present, observable competence to perform a learned psychomotor act" (section 1607.16.T). In contrast, AO requirements are couched in terms of much more abstract *hypothetical constructs*. To quote the APA *Standards* (1985), in this context a construct is

> a psychological characteristic (e.g., numerical ability, spatial ability, introversion, anxiety) considered to vary or differ across individuals. A construct

Summary

Although the regulatory climate in which job analysis data are collected and applied is constantly changing, some general principles have remained relatively constant. First, behaviorally specific job analysis is the only effective strategy for identifying and (content) validating KS-based job specifications; demonstrable links to such data are also vital for developing and defending criterion measurement and appraisal systems.

Second, the techniques for identifying and validating AO-based job specifications are qualitatively distinct from those used to validate KS-based requirements. Although many types of job analysis data can be used in the process of (construct) validating inferences of general ability and trait requirements, specifications of ability requirements must be based on something more than just the "professional judgment" of the analyst.

Third, a job analysis is needed to justify validity generalization or transportability decisions, although the specificity of this information is open to debate. To the extent that behaviorally specific predictor tests are used (e.g., work samples), it is likely that more specific job analysis data should be used to assess job similarity.

Uses of Job Analysis Information

Job analysis data form—or at least *should* form—the foundation for nearly all important personnel decisions. This section will examine the link between job analysis data and the uses to which it may be put. On the positive side, a number of the links between job analysis data and specific personnel functions are quite clear and effective, such as developing performance appraisal instruments and predicting compensation rates; conversely, some of these links are highly controversial and in need of significant additional research before clear guidelines can be given to practitioners, such as in using job analysis data to set ability- and trait-based worker specifications for employee selection purposes.

Job Classification

Perhaps the most basic and direct use of job analysis information is to determine what constitutes a *job* or a *job family*. That is, which groupings of positions or jobs are similar enough to be treated interchangeably for a specified personnel purpose? Job classification decisions have become increasingly important as a consequence of recent interest in validity generalization techniques; specifically, to justify the generalizability of previous validation results, one must be able to show that new jobs are sufficiently similar to the ones on which validation data exist. Several distinct steps are involved in arriving at such a determination.

Theoretical Models. As a first step in making a job classification decision, one must decide which conceptual view of jobs and job families is to be taken. Until recently, this decision was implicit, as only one view of the underlying structure of work was commonly encountered in personnel psychology.

The traditional view of work is based on what numerical taxonomists and factor analysts have termed an *independent-cluster structure* (e.g., see Coombs & Satter, 1949; Harris & Kaiser, 1964; Sokal, 1974). In independent-cluster structures, each of the entities to be grouped is classified into one *and only one* class or grouping (numerical taxonomists refer to these as *taxons*). In the context of job classification, this means that when positions are grouped to form jobs, each position will be assigned to one and only one job title grouping, and there will be no overlap between titles. Similarly, when grouping jobs to form job families, each job will be placed into one and only one job family cluster, and no overlap will exist between job families. A pictoral view of these links is presented in Figure 5.

I have questioned this conceptual view of work (e.g., Harvey, 1982, 1986a); specifically, when grouping jobs, one often encounters patterns of job similarity that simply do not behave according to these simple rules. When grouping jobs to form job families, instead of finding independent-cluster structures, the data often indicate that family membership can best be represented as *overlapping* clusters of jobs. That is, some jobs belong to multiple families. Similarly, when grouping positions to form jobs, positions often are classified into multiple jobs; put another way, there is overlap between the job clusters.

Researchers in the field of numerical taxonomy have long been aware of such phenomena; indeed, some (e.g., Sokal, 1974) have concluded that overlapping-cluster structures frequently provide a more realistic representation of the similarity of objects than the simplistic independent-clusters model. For similar reasons, I contend that the traditional independent-cluster model is usually a fundamental oversimplification of the interrelations between positions, jobs, and job families. For example, in the case of grouping jobs to form families, only in cases in which there is effectively no overlap between jobs in terms of having important job activities in common can independent-cluster job families be produced; such findings are empirically unlikely, especially when worker-oriented items or job dimensions are used to assess job similarity.

Under the overlapping-clusters view of job similarity, jobs will (a) vary with respect to a number of general dimensions of work activity, (b) be similar to one another in requiring some dimensions and dissimilar regarding others, and (c) typically each involve *a number* of general dimensions of work activity. For example, one job family might be characterized by dimensions like external contacts, supervising nonsupervisory employees, and operating stationary machines, including jobs like loading dock supervisor—whereas a second family might be composed of jobs involving external contacts, operating light highway vehicles, hazardous working conditions, supervising nonemployees, and supervising nonsupervisory employees, including such jobs as police sergeant and fire lieutenant. These two families overlap in terms of both having external contacts and supervising nonsupervisory employees, but are nonoverlapping on their other important dimensions.

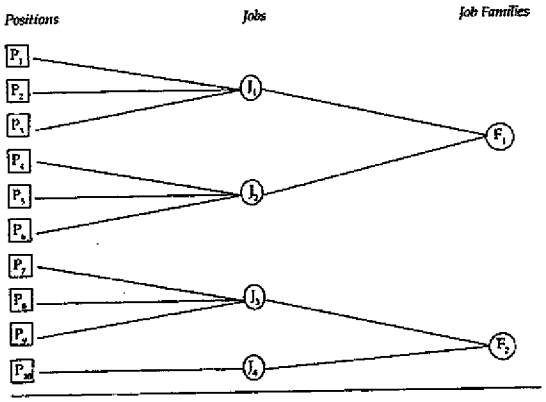
In essence, this is the conceptual view of work that underlies the *job component validity* technique (e.g., McCormick, DeNisi, & Shaw, 1979). That is, even highly task-heterogeneous jobs can be found similar in terms of sharing one or more general dimensions of work, and dissimilar on other general dimensions of work.

Techniques for Making Job Classification Decisions. A second decision that must be made when grouping jobs is to select a grouping procedure. This involves selecting from among the numerous quantitative techniques for analyzing multivariate data that have been proposed as vehicles for identifying job similarity. Although some authors have claimed that holistic job grouping judgments (e.g., Pearlman, 1980) can be used to make such decisions, I would argue that the limitations of holistic techniques—in particular, the job-relatedness issue and the lack of adequate demonstrations of results-similarity between holistic and empirical job grouping methods—are at present sufficient to remove them from consideration.

With regard to selecting a statistical job grouping technique, the descriptive, dimension-oriented techniques (e.g., Q-factor analysis, three-mode factor analysis) offer significant advantages over both inferential methods (e.g., MANOVA) and hierarchical cluster analysis (HCA; see Harvey, 1986a, for further discussion). The superiority of dimension-oriented techniques is enhanced to the degree that the overlapping-clusters

FIGURE 5

Traditional View of the Link Between Positions, Job, and Job Families



model of job similarity is more appropriate for the data than the independent-clusters model. Regarding the inferential job-grouping methods, limitations include (a) the critical role of statistical power (i.e., with a large enough sample of job descriptors or raters, virtually *any* pair of jobs can be found to be "significantly different" because of the power to detect even trivial differences; with a small enough sample, any set of jobs can be found to be "the same" due to lack of power to detect even large true differences); (b) the lack of effect size guidelines to guide decision making when significance tests are not used; and (c) testing a null hypothesis that few would ever expect to be true (i.e., if measured with sufficient precision, it is unlikely that two jobs would ever demonstrate *exactly equal* job analysis profiles).

Limitations of hierarchical cluster analysis include (a) HCA can *only* produce independent-cluster solutions, regardless of the actual latent structure of the data; (b) the decision rules for identifying the correct number of clusters in HCA have never been definitive and are almost certainly going to be incorrect when the independent-clusters assumption is violated (i.e., if the jobs truly do not form independent clusters, the concept of a "correct" number of independent clusters is meaningless); and (c) numerous algorithms for computing HCA solutions exist, and different computational methods often produce very different cluster solutions (e.g., Milligan, 1981). In short, there is little point in forcing an overly simplistic conceptual model onto a set of job similarity data. Indeed, if independent clusters *are* actually present, dimension-oriented techniques like Q-factor analysis are capable of recovering them; the reverse is not true of hierarchical cluster analysis (i.e., if overlapping clusters are present, it is highly unlikely that HCA could recover them).

The important question to be addressed when grouping positions or jobs is *on which attributes* do they differ and on which attributes are they similar—not are they the "same" or "different." Particularly when grouping jobs from different plants, geographical areas, or industries (e.g., to conduct a consortium-style test validation study), it is inevitable that *some* degree of difference on some job dimensions will be observed. What is needed is research that will determine the amount and kinds of differences that can be tolerated before validities or other personnel decisions are moderated by cross-job behavior differences. Although initial efforts toward this goal have been promising (e.g., Gutenberg, Arvey, Osburn, & Jeanneret, 1983), much additional research remains to be conducted.

Selecting the Job Analysis Data. A third decision that must be made when conducting job classification analyses is to select the appropriate job analysis philosophy when describing the positions or jobs to be grouped. Although only a few studies have varied the type of job analysis data to assess the degree to which job classification structures differ (e.g., Cornelius, Carron, & Collins, 1979; Harvey, 1988), commonsense and the available empirical data indicate that the kind of job descriptors chosen will exert a powerful effect on the resulting job groupings. As Harvey (1988) reported, more behaviorally and technologically abstract descriptors (in this case, JEI versus task-oriented data) produced more abstract and overlapping job families. Consistent with the conceptual foundation of the worker-oriented approach, its descriptors are less sensitive to task differences between jobs, and job groupings identified using worker-oriented data will be based on similarity in terms of more generalized work behaviors.

Even when holding the job analysis philosophy constant, different job similarity statistics—as well as ways of centering and/or standardizing the data matrix—will produce different job classification conclusions (e.g., Hamer & Cunningham, 1981; Harvey, 1985a; Harvey & Wilson, 1985). Some statistics are based solely on dichotomous similarity (e.g., percent tasks in common), some equate individual job profiles for different total rating scores prior to computing similarity (e.g., the CODAP percent-task-overlap statistic), some are sensitive only to similarity in profile shape and eliminate profile elevation differences (e.g., correlations), and others retain information on shape, elevation, and dispersion differences (e.g., profile distance statistics).

Given that different types of job analysis data and different methods for computing similarity statistics will yield different solutions, it is important to match the purpose for grouping jobs with the job analysis philosophy. For applications in which a high degree of technological detail and homogeneity are desired (e.g., training, performance appraisal, job title assignment, developing job descriptions, and other functions that use the *job* as the basic unit of analysis), job grouping analyses should be conducted using more specific, task-oriented job descriptors as input. Conversely, for purposes that desire groupings that are similar in terms of generalized work behaviors (e.g., job component validity, test validation using general ability tests, career path planning), worker-oriented items or job dimension scores would be preferred. The choice of the profile similarity statistic will be most strongly influenced by the degree to which cross-job differences in profile elevation are meaningful: If within-job-relative rating scales are used to rate the job analysis items, statistics sensitive to either binary profile similarity (e.g., percent items in common) or

relative profile shape (but not elevation) should be used (e.g., profile correlations). In either case, interpretations of the job groupings must be made in the context of the lack of meaningful elevation information (i.e., jobs that group together may differ significantly in the actual levels of the job behaviors involved, which may or may not be appropriate). If cross-job level comparisons are interpretable (e.g., Type 2 or 3 job analysis data are used as input), elevation-sensitive statistics may be used instead and will provide a more comprehensive description of job similarity.

Job Descriptions

Another direct use to which job analysis information may be put is the formation of job descriptions. However, what is meant by *job description* is quite variable; for example, the popular *narrative job description* is often only loosely based on a systematic job analysis. Figure 6 presents an example of a narrative job description.

Job descriptions need not be strictly narrative; Figure 7 presents an example of a job description developed from a task-oriented job analysis data base; computerized integrated personnel systems are capable of automatically generating such descriptions from current personnel data and updating the descriptions as job tasks and duties change. The increased job relatedness and specificity of job descriptions of this sort are powerful inducements to prefer this approach over subjectively developed job narratives.

Performance Appraisal

As noted earlier, the courts have for some time required employers to demonstrate the job relatedness of appraisal systems (e.g., *Brito v. Zia*, 1973; *Watson v. Fort Worth Bank & Trust*, 1988). In practical terms, demonstrating the job relatedness of an appraisal system means showing that a job analysis was performed and identifying the link between the job analysis data and the appraisal procedures.

An assessment of whether two key errors are present should also occur. *Criterion contamination* is said to occur when the appraisal procedure involves the rating of factors that are not demonstrably part of the job (e.g., Cascio, 1987). For example, rating an employee on the dimension "dealing with the public" is only appropriate if the job actually requires the employee to engage in public contact. *Criterion deficiency* occurs when the appraisal system fails to rate aspects of performance that *are* part of the job. Both errors concern the content validity of the appraisal system, the former involving rating of dimensions that are not in the domain of job activities, the latter involving inadequate domain sampling.

Using job analysis information in the performance appraisal process would seem to be a very straightforward issue. However, a number of different strategies for linking job analysis data to the appraisal process exist, and they vary in terms of the strength of this link.

Holistic Methods. Schmidt et al. (1981) concluded that individual effectiveness on the job must be defined as a single holistic score and that ratings on a number of behaviorally specific performance dimensions are unnecessary and potentially misleading. The basis for this prescription is the claim that:

> correlations between criterion dimensions, after correction for attenuation due to unreliability, typically approach 1.00, indicating that different behavioral dimensions are virtually collinear at the true score level (Schmidt & Hunter, 1978). Under these circumstances, it is obvious that only a measure of overall job performance is needed in validity studies. (Schmidt et al., 1981, p. 175)

FIGURE 6

Narrative Job Description

City Architect I

Nature of Work

This is professional and technical work in the preparation of architectral plans, designs, and specifications for a variety of municipal or public works building projects and facilities.

Minimum Qualifications

Education and Experience—Graduation from an accredited college or university with a specialization in architecture or architectural engineering; or equal.

Knowledge, Abilities, and Skills—Considerable knowledge of the principles and practices of architecture; ability to make structural and related mathematical computations and make recommendations on architectural problems; ability to design moderately difficult architectural projects; ability to interpret local building codes and zoning regulations; ability to secure good working relationships with private contractors and employees; ability to train and supervise the work of technical and other subordinates in a manner conducive to full performance; ability to express ideas clearly and concisely, orally and in writing; skill in the use of architectural instruments and equipment.

Illustration of Duties

Prepares or assists in the preparation of architectural plans and designs all types of building projects constructed by the City, including fire stations, park and recreation buildings, office buildings, warehouses, and similar structures; prepares or supervises the preparation of final working drawings including architectural drawings, such as site plans, foundations, floor plans, elevations, section details, diagrams, and schedules rendering general features and scale details; prepares or supervises some of the engineering calculations, drawings and plans for mechanical details, such as plumbing, air-conditioning phases, and lighting features; writes construction standards and project specifications; prepares sketches including plans, elevations, site plans, and renderings and makes reports on feasibility and cost for proposed City work; writes specifications for all aspects of architectural projects including structural, mechanical, electrical, and air-conditioning work; confers with engineering personnel engaged in the preparation of structural plans for a building, making recommendations and suggestions as to materials, construction, and necessary adjustments in architectural designs to fit structural requirements; inspects construction in the field by checking for conformity with plans and material specifications; inspects existing structures to determine need for alterations or improvements and prepares drawings for such changes; performs related work as required.

Supervision Received

General and specific assignments are received and work is performed according to prescribed methods and procedures with allowance for some independence in judgment in accomplishing the assignments.

Supervision Exercised

Usually limited to supervision of technical assistants in any phase.

FIGURE 7

Task-based Job Description

Job Description Profile Scores for TM Pay Grade E-7 D-Tsk	Task Title	*Percent Members Performing*	*Avg % Time Spent by Members Performing*	*Avg % Time Spent by All Members*	*Cum. Sum of Avg % Time Spent by All Members*	*No. Duties or Tasks*
B 1	Write enlisted performance evaluations	80.00	2.46	1.97	1.97	
A 1	Review enlisted performance evaluations	82.86	2.32	1.93	3.90	
C 18	Maintain logs (pass down log [PDL] etc.)	62.86	3.01	1.89	5.79	
B 4	Ensure work assigned to subordinates is completed	82.86	2.27	1.88	7.66	
C 6	Update publications/instructions (pen and ink and page changes)	80.00	2.28	1.82	9.49	5
B 5	Coordinate work within division	68.57	2.56	1.76	11.25	
C 17	Fill out work requests/work orders	62.86	2.42	1.52	12.77	
A 11	Evaluate operational commitments in order to schedule workload	60.00	2.35	1.41	14.18	
A 5	Screen messages, bulletins, etc. for appropriate action	74.28	1.84	1.36	15.54	
A 2	Make personnel assignments	80.00	1.68	1.34	16.88	10
A 3	Assign work priorities	71.43	1.78	1.27	18.15	
C 7	Maintain correspondence/message files	62.86	1.94	1.22	19.38	
A 24	Receipt for weapons	71.43	1.68	1.20	20.58	
B 2	Make work assignments	74.28	1.56	1.16	21.74	
A 25	Ensure readiness of command for inspections (administrative operational, material, etc.)	68.57	1.66	1.14	22.88	15
A 15	Prepare weekly discrepancy report	54.28	2.01	1.09	23.97	
C 25	Prepare reports of unsatisfactory/defective torpedoes, or equipment	65.71	1.64	1.08	25.05	
A 27	Review and submit status reports (performance, inventory, casualty, etc.)	62.86	1.64	1.03	26.08	
F 21	Inspect weapons handling gear (slings, hoist, etc.)	68.57	1.37	0.94	27.02	
C 5	Maintain tickler file	48.57	1.88	0.91	27.93	20
Z 9	Attend meetings, seminars, conferences, etc.	54.28	1.68	0.91	28.84	
A 9	Monitor training program	57.14	1.54	0.88	29.72	
E 12	Inspect all material upon receipt for damage, quality, quantity, etc.	42.86	2.02	0.87	30.58	
Z 4	Stand inspections	60.00	1.42	0.85	31.44	
C 11	Route correspondence/publications/instructions, etc.	51.43	1.65	0.85	32.28	25

FIGURE 7

Task-based Job Description (continued)

Job Description Profile Scores for TM Pay Grade E-7 D-Tsk	Task Title	*Percent Members Performing*	*Avg % Time Spent by Members Performing*	*Avg % Time Spent by All Members*	*Cum. Sum of Avg % Time Spent by All Members*	*No. Duties or Tasks*
A 18	Coordinate weapon overhaul and repair within own command and/or between other ships and stations	51.43	1.64	0.84	33.12	
A 20	Recommend personnel for formal training	65.71	1.26	0.83	33.95	
D 11	Sign off practical factors	71.43	1.14	0.81	34.76	
F 32	Turn in torque wrenches for calibration	51.43	1.55	0.80	35.56	
C 39	Prepare/update 3M schedules (cycle, quarterly, weekly)	54.28	1.42	0.77	36.33	30
C 1	Draft naval messages	48.57	1.58	0.76	37.09	
A 14	Sign requisitions requiring approval	54.28	1.39	0.76	37.85	
C 8	Type correspondence/forms	40.00	1.82	0.73	38.58	
A 10	Represent command at conferences and meetings	45.71	1.56	0.71	39.29	
Z 1	Hold field days, sweepdowns, etc.	42.86	1.60	0.68	39.97	35
C 20	Update recall bill	37.14	1.84	0.68	40.66	
Z 6	Counsel personnel on personal/military matters	54.28	1.25	0.68	41.34	
D 2	Update individual training records	45.71	1.48	0.68	42.01	
D 1	Prepare individual training records	45.71	1.48	0.68	42.69	
C 35	Maintain torpedo record book	51.43	1.32	0.68	43.36	40
A 7	Coordinate with military activities as required	48.57	1.37	0.66	44.03	
F 62	Destroy classified materials in accordance with current instructions	51.43	1.29	0.66	44.69	
C 15	Draft instructions/notices	45.71	1.45	0.66	45.35	
D 3	Schedule training lectures	54.28	1.22	0.66	46.01	
C 24	Maintain log/file of report of unsatisfactory/defective torpedoes or equipment	57.14	1.14	0.65	46.66	45
C 16	Review/chop outgoing correspondence/messages	34.28	1.89	0.64	47.31	
A 22	Maintain liaison with personnel of other departments to prevent or correct interface problems	45.71	1.41	0.64	47.95	
B 7	Complete weapons firing reports	45.71	1.40	0.64	48.58	
C 13	Maintain status boards	57.14	1.10	0.63	49.21	
Z 5	Attend general drills	48.57	1.28	0.62	49.83	50
A 19	Determine expendable materials (surveys, disposal, etc.)	51.43	1.19	0.61	50.44	
A 16	Evaluate and take appropriate action on reports from torpedo readiness acceptance (TRAT) inspection	51.43	1.16	0.60	51.04	
F 53	Perform weapons receipt inspection	54.28	1.09	0.59	51.63	
A 4	Write billet/job descriptions	45.71	1.28	0.58	52.22	
E 17	Pack/unpack weapons/components	48.57	1.20	0.58	52.80	55

FIGURE 7

Task-based Job Description (continued)

Job Description Profile Scores for TM Pay Grade E-7			*Percent Members Performing*	*Avg % Time Spent by Members Performing*	*Avg % Time Spent by All Members*	*Cum. Sum of Avg % Time Spent by All Members*	*No. Duties or Tasks*
D-Tsk		*Task Title*					
J	41	Perform quality assurance checks on weapons	17.14	3.37	0.58	53.37	
C	19	Maintain leave schedules	45.71	1.23	0.56	53.94	
C	29	Make entries in daily work log	40.00	1.40	0.56	54.50	
F	61	Participate in weapons firefighting procedures	45.71	1.22	0.56	55.06	
C	14	Distribute safety material (publications, posters, etc).	51.43	1.06	0.55	55.60	60
E	1	Order parts, tools, supplies, etc.	57.14	0.95	0.54	56.14	
C	12	Maintain division officer's notebook	31.43	1.67	0.52	56.67	
A	23	Organize departmental/division security	42.86	1.20	0.52	57.18	
F	1	Test weapons security alarm systems	51.43	1.00	0.51	57.70	
C	2	Draft naval letters	34.28	1.46	0.50	58.20	65
F	44	Remove/install weapons/components in shipping containers	42.86	1.16	0.49	58.69	
D	22	Develop on-the-job training (OJT) program	45.71	1.07	0.49	59.18	
Z	8	Conduct inspections (zone,personnel, safety, etc.)	48.57	1.01	0.49	59.66	
F	8	Chip, preserve, and paint topside areas	11.43	0.76	0.08	95.28	
F	24	Operate forklift	5.71	1.50	0.08	95.37	
H	17	Install battery power supplies	14.28	0.60	0.08	95.45	
P	7	Handle and fire pyrotechnic devices	17.14	0.50	0.08	95.53	
D	20	Prepare and administer feedback reports for the purpose of updating training	11.43	0.71	0.08	95.61	230
C	26	Prepare corrective action request (NAVORD form 4855/18)	11.43	0.70	0.08	95.69	
F	25	Maintain/use hydraulic RAM	11.43	0.70	0.08	95.77	
G	22	Electrically zero synchros/servos	8.57	0.98	0.08	95.85	
G	10	Remove/replace components on printed circuit boards	8.57	0.96	0.08	95.93	
G	9	Repair cables (splices, etc.)	8.57	0.94	0.08	96.01	235
F	14	Perform emergency de-fueling procedures on weapons	11.43	0.67	0.08	96.09	
J	30	Overhaul and repair pneumatic actuated valves	5.71	1.38	0.08	96.17	
C	32	Prepare work request customer service form (OPNAV form 4790/36A)	8.57	0.88	0.07	96.24	
D	10	Prepare test/examinations	11.43	0.64	0.07	96.31	
D	16	Construct training aids	11.43	0.64	0.07	96.38	240

FIGURE 7

Task-based Job Description (continued)

Job Description Profile Scores for TM Pay Grade E-7			*Percent Members Performing*	*Avg % Time Spent by Members Performing*	*Avg % Time Spent by All Members*	*Cum. Sum of Avg % Time Spent by All Members*	*No. Duties or Tasks*
D-Tsk		*Task Title*					
F	5	Maintain small armsiclean, lubricate, etc.)	14.28	0.51	0.07	96.45	
J	18	Test weapons homing control logic unit	5.71	1.25	0.07	96.52	
J	23	Test weapons velocity switches	11.43	0.62	0.07	96.66	
H	33	Maintain torpedo tube electrical system	11.43	0.61	0.07	96.73	245
F	33	Calibrate torque wrenches	14.28	0.50	0.07	96.79	
J	57	Test missile igniter	8.57	0.78	0.06	96.86	
C	37	Review/update casualty reports (CASREPTS)	8.57	0.78	0.06	96.92	
F	42	Operate forward/aft capstan	11.43	0.57	0.06	96.99	
J	10	Service weapons hydraulic systems	5.71	1.16	0.06	97.05	250
J	43	Fuel/defuel weapons	14.28	0.46	0.06	97.11	
D	21	Prepare programmed instructions	8.57	0.74	0.06	97.17	
F	48	Inspect/test-operate magazine de-watering systems	8.57	0.70	0.06	97.23	
H	11	Perform final preparation of complete torpedo (MK-16)	17.14	0.35	0.06	97.29	
D	25	Train instructors in OJT methods	8.57	0.66	0.06	97.35	255
J	31	Overhaul and repair mechanical depth mechanisms	2.86	2.00	0.06	97.41	
J	17	Remove/replace weapons propulsion battery	8.57	0.70	0.06	97.46	
J	14	Overhaul and repair weapon turbine propulsion unit	2.86	2.00	0.06	97.52	
F	79	Install safety wire	8.57	0.69	0.06	97.58	
J	33	Overhaul and repair mechanical steering units	2.86	2.00	0.06	97.63	260
J	20	Repair weapons gyros	2.86	2.00	0.06	97.69	
H	19	Remove/replace thrust reversal nozzle plug	11.43	0.50	0.06	97.74	
H	13	Perform abort procedures on weapons	14.28	0.42	0.06	97.80	
F	78	Neutralize electrolyte spillage (acid, alkaline)	8.57	0.65	0.05	97.85	
Z	14	Stand special sea detail watches (helmsman, after steering, line handler, etc.)	8.57	0.62	0.05	97,90	265
F	65	Clean/repair liquid stowage tanks	5.71	0.88	0.05	97,95	
J	49	Install electrolyte in weapons batteries	8.57	0.60	0.05	98.00	
J	9	Perform torpedo receiver sensitive test	5.71	0.86	0.05	98.05	

From *Methods to Evaluate Scales and Sample Size for Stable Task Inventory Information* by J. Fass and O. Robertson, 1980, San Diego, CA: Navy Personnel Research and Development Center (NPRDC TR No. 80-28). Reprinted by permission.

There is reason to be skeptical of the generality of this conclusion. In a test of the Schmidt-Hunter unidimensionality hypothesis, Butler and Harvey (1986) examined the corrected and raw correlations between the dimensions contained in a behaviorally specific appraisal system developed for police patrol officers. Although the dimensions were correlated to some degree, there was no evidence of the high levels of redundancy claimed by Schmidt et al. (1981).

It seems reasonable to predict that rating dimension correlations will be lower to the extent that the appraisal dimensions—and the rating points on each dimension—are defined in terms of observable job behaviors (e.g., see Feldman, 1981); conversely, high correlations may result when vague traits are rated. Thus, when a traditionally specific job analysis is conducted and behaviorally explicit dimensions are developed, the problem of excessive correlation cited by Schmidt et al. (1981) as justification for using a single global effectiveness rating should not occur. Unfortunately, few research studies conducted in realistic settings have addressed this question.

Trait-oriented Methods. The most salient characteristic of the trait approach is that a number of worker traits judged necessary for successful job performance are listed, and the amount of each trait possessed by the worker is judged. Worker performance is deemed more effective to the degree that employees possess higher levels of the traits. Figure 8 presents an instrument based on this approach.

The main problems with trait-oriented appraisal instruments concern criterion contamination and deficiency. Basically, in the absence of a job analysis that defines the domain of important job behaviors, it is impossible to demonstrate that a trait-oriented instrument spans the domain of job performance fully and does not include extraneous dimensions. Even if a job analysis has been performed, the process by which the developer accomplishes the inferential leap from job analysis to a set of trait listings must be fully documented; depending on the traits chosen, this may not be a trivial task.

Task-oriented Methods. Task-oriented approaches to performance appraisal differ from the trait-based methods primarily in terms of the way in which performance dimensions are identified. In trait systems, job descriptive data are used to infer the general traits that characterize successful job performance in much the same way as job specifications are developed for selection purposes. Task-based systems employ a much more direct link to the job analysis data (e.g., Harvey, 1986b; Thompson & Thompson, 1985).

Once the task-based job analysis is completed and duty categories have been indentified, a task-based appraisal instrument can be constructed by using the duties as the appraisal dimensions and using the tasks performed under each duty to define the dimension. For administrative purposes, only duty ratings need be collected; for purposes of employee development, the job tasks themselves may also be rated and these ratings summarized for each dimension. A portion of a task-based developmental appraisal form for the job of computer programmer is presented in Figure 9. Advantages of this approach to using job analysis data for performance appraisal center on the ease of demonstrating the job relatedness of the rating forms as well as being able to justify dimension ratings in terms of actual job behaviors.

Critical Incident Methods. There are three primary means by which critical incident data may be used in the appraisal process: (a) behaviorally anchored rating scales (BARS; e.g., Smith & Kendall, 1963) and their derivatives, (b) weighted behavior checklists, and (c) a hybrid "task-BARS" approach involving the use of critical incidents to define performance

FIGURE 8

Trait-based Appraisal Instrument

Worker Requirements Study Ratings Sheet

Circle Appropriate Number

1. Quality of Work (accuracy, neatness, thoroughness)

Inferior Work			Rather Careless			Meets Requirements			Highly Accurate			Exceptional		
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

2. Quantity of Work (volume, amount, speed)

Very Slow			Insufficient Work			Moderate			Rapid Worker			Highly Productive			
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

3. Knowledge of Work

Almost None		Limited		Adequate		Good Understanding		Excellent Comprehension	
0	1	2	3	4	5	6	7	8	9	10

4. Adaptability (adjustment to change, ability to learn)

Unable To Adapt		Slow in Learning		Satisfactory		Adapts Readily		Rapid Learner	
0	1	2	3	4	5	6	7	8	9	10

5. Dependability (reliability)

Needs Constant Supervision		Needs Frequent Checking		Usually Dependable		Seldom Needs Checking		Highly Reliable	
0	1	2	3	4	5	6	7	8	9	10

6. Cooperation (working with other employees)

Troublemaker		Has Difficulty		Generally Cooperative		Gets Along Well		Excellent Relations	
0	1	2	3	4	5	6	7	8	9	10

From Establishing Valid and Fair Worker Requirements: Claims Examiner I (Project Report, p. 39) by R. Klimoski and S. Gravenkemper, 1982, Columbus, OH: Bureau of Disability Determination. Copyright 1982 by Bureau of Disability Determination. Reprinted by permission.

standards in a task-oriented appraisal instrument (e.g., as was used by Butler & Harvey, 1986). In the BARS approach, a pool of critical incidents is generated, the mean effectiveness rating of each incident is determined, incidents are grouped by SMEs into clusters to form the appraisal dimensions, and a subset of the scaled incidents is used to define the various levels of performance for each dimension. Figure 2 presents a typical BARS dimension rating scale.

The *BARS approach* is useful in that the appraisal dimensions are described in explicit behavioral terms. Unfortunately, it has several drawbacks. First, gaps may exist in the rating scales for each dimension (i.e., there are usually areas in which no critical incidents have scale values); thus, raters must interpolate between incidents. Second, there is often considerable task heterogeneity within BARS dimensions regarding critical incident content; that is, for a single dimension, the critical incidents used as anchors may describe the performance of very different job tasks (e.g., in Figure 2, the most effective incident deals with providing training to new employees, an intermediate incident deals with criticizing store standards, and the least effective incident deals with lying to employees about salary policies). It is entirely possible that a single employee could demonstrate *all* of these behaviors during the rating period, as they exemplify good and bad ways of performing what are basically *different tasks;* which incident should be chosen as the most representative? Conversely, what if the employee has performed *none* of the incidents defining the dimension?

Third, questions about the job relatedness of BARS can be raised, due to the fact that the BARS method *discards* critical incidents at nearly every step of the development process—for example, if SMEs can't agree on the effectiveness rating, if SMEs can't agree on the performance dimension measured by the incident, or if several incidents with the same scale value exist for a given dimension. Systematically ignoring parts of the job analysis is the operational definition of criterion deficiency.

The second critical incident method—*weighted checklists*—avoids these problems by presenting raters with the entire list of critical incidents generated from the job analysis. By weighting each checked behavioral incident by its scale value, numerical dimension ratings can be obtained. However, the degree to which the pool of critical incidents spans the domain of job tasks can be questioned for both the incident checklist and BARS approaches; that is, if the job analysis conducted in the course of appraisal instrument development consists solely of the generation of critical incidents, it is possible that job tasks exist for which there are no corresponding critical incidents. Again, potentially serious criterion deficiency could exist.

The third approach—what I have termed the *task-BARS technique*—is a hybrid combining features of both the task-based and BARS techniques, using critical incidents as rating anchors in the general context of a task-based appraisal format. In short, instead of using a Likert-type rating scale to rate employee effectiveness on the tasks or dimensions (e.g., with anchors like *Far below expectations, Below expectations, At the expected level, Above expectations, Far above expectations),* appraisal dimension anchors are composed of critical incidents representing *all possible levels* of dimension/task performance. Thus, critical incidents serve as the behavioral anchors for task-identified rating dimensions. Because the task-based rating dimensions are generated first, and critical incidents are formed to define each of the various possible levels of task effectiveness, the problems with gaps in the rating scales and task heterogeneity within dimensions seen in traditional BARS scales can be reduced. A sample dimension from the instrument used by Butler and Harvey (1986) is presented in Figure 10.

FIGURE 9

Task-based Performance Appraisal Dimension

Duty Category	*Does Not Apply*	*Performance on This Task*	
		Needs Improvement	*Is Acceptable*
Writing, Completing Reports			
Complete standard forms for city, state, federal authorities	____	____	____
Draft written evaluations of subordinate's performance	____	____	____
Complete monthly activity report	____	____	____
Write letters, memos	____	____	____
Write specifications, plans	____	____	____
Complete request form for checks	____	____	____
Complete suspect booking, custody, control forms, contact, and fingerprint cards	____	____	____
Prepare service requests, job orders	____	____	____
Make written record of information supplied to police officers	____	____	____
File charges with appropriate agency	____	____	____
Prepare monthly, quarterly, annual reports	____	____	____
Prepare purchase requisitions	____	____	____
Prepare service requests for vehicle repairs	____	____	____
Complete offense reports	____	____	____
Complete accident reports	____	____	____
Complete vehicle maintenance checklists	____	____	____
Prepare reports for supervisor	____	____	____
Recordkeeping			
Process requests for vehicle repair	____	____	____
Maintain records of maintenance needs on equipment	____	____	____
Record mileage and amount of gas used	____	____	____
Maintain records of attendance, time off, sick, vacation leaves	____	____	____
Input data into computer	____	____	____
Record arrest information	____	____	____
Record statements of witnesses, suspects, victims	____	____	____
Record arrival of units at scene	____	____	____
Receive and record alarms	____	____	____
Log types and kinds of evidence	____	____	____
Record types of injuries suffered by victims	____	____	____
Note sighted water leaks	____	____	____
Keep notes on employee performance for evaluation purposes	____	____	____
Record daily activities	____	____	____
Record work, hours on time sheet	____	____	____
Maintain daily contact sheet	____	____	____
Complete vehicle check list	____	____	____
Process unpaid violations	____	____	____
Maintain records of vehicle maintenance needs	____	____	____
Videotape bookings	____	____	____
Record sworn affidavits	____	____	____

FIGURE 10

Task-BARS Rating Dimension

Patrol Officer

Preparing for Duty

(1) Late for roll call majority of period, does not check equipment or vehicle, does not have necessary equipment to go to work.

(2) Late for roll call, does not check equipment or vehicle for damage or needed repairs, unable to go to work from roll call, has to go to locker, vehicle, or home to get necessary equipment.

(3) Not fully dressed for roll call, does not have all necessary equipment.

(4) On time, has all necessary equipment to go to work, fully dressed.

(5) Early for work, has all necessary equipment to go to work, fully dressed.

(6) Always early for work, gathers all necessary equipment to go to work, fully dressed, checks activity from previous shifts before going to roll call.

(7) Always early for work, gathers all necessary equipment to go to work, fully dressed, uses time before roll call to review previous shift's activities and any new bulletins, takes notes of previous shift's activity mentioned during roll call.

In summary, performance appraisal and criterion construction methods vary considerably in terms of the ease with which they may be demonstrated to be job related. For both legal defensibility and sound personnel management reasons, techniques that are based on specific, observable job behaviors (i.e., the task and critical incident methods) should be easier to use and defend than methods having only a weak or indirect link to a job analysis (e.g., trait-based and holistic methods).

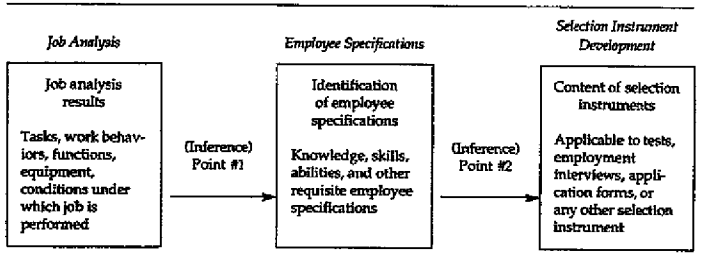
Employee Selection

Identifying KS and AO Requirements. In traditional approaches to test validation, employers are faced with the need to make two separate kinds of inferences when linking job analysis data to the employee selection process. First, the behavioral descriptions of job activities obtained from the job analysis must be converted to a listing of required KS and AO requirements; this is the process of developing job specifications examined earlier. Second, the KS/AO specifications must be tied to the practical processes by which job candidates are assessed with respect to their levels of the necessary KS and AO constructs (i.e., a test development or employee assessment question). Figure 11 diagrams these processes.

Identifying Required KSs and AOs. When developing KS/AO specifications, the courts and most professional standards documents require a detailed analysis of important job activities, regardless of whether a content- or construct-oriented strategy is followed. When task-oriented data are used, the first problem is to identify a job's *critical* or *important* tasks.

FIGURE 11

Inferential Steps in Using Job Analysis Data to Develop Employee Selection Techniques



Combining task-oriented data with critical incidents information can also be useful in identifying critical tasks (e.g., *Davis v. Washington*, 1972). Once critical job behaviors have been identified, SMEs can be used to identify the KSs and AOs required to perform the critical tasks—ideally, with as much direct reference to specific job tasks as possible, especially for the KS items—and assess their relative importance as determinants of job success (e.g., see *Contreras v. City of Los Angeles*, 1981; *Guardians Association of NYC Police Dept. v. Civil Service Commission of New York*, 1980).

If a worker-oriented job analysis has been conducted, one may instead rely on empirical links between worker-oriented scores and KS/AO requirements identified previously to predict KS/AO requirements for the job at hand (e.g., using the job component validity approach developed by McCormick, DeNisi, & Shaw, 1979; McCormick et al., 1972). Once sufficient empirical data exist to link the dimensions of work behavior (identified via worker-oriented instruments) with the dimensions of human knowledge, ability, and skill (identified via standardized ability and achievement tests, as appropriate), the empirical-linking approach will undoubtedly be preferred over the potentially highly subjective processes involved in using SMEs to infer job specifications (e.g., see Dunnette & Borman, 1979, pp. 484–485).

With respect to the use of SMEs to identify KS and AO requirements, considerable variability in the strategies used to link job analysis data to KS/AO inferences exists. For example, *threshold traits analysis* (TTA; Lopez, Kesselman, & Lopez, 1981) uses job supervisors to directly rate the skills, job knowledge, specific abilities, abstract abilities, and personality traits they feel are required to perform the job. A listing of the traits rated in TTA is presented in Figure 12.

To support the claim that TTA ratings of KS and AO constructs are job-related, Lopez et al. noted that in one study they conducted, the median interrater reliability computed across the 33 traits was .86 (however, it is not clear

FIGURE 12

KS and AO Traits Rated in Threshold Traits Analysis

	Area	Job Functions	Trait	Description—Can:
Threshold Traits Analysis	*Physical*	Physical exertion	1. Strength	Lift, pull, or push physical objects
			2. Stamina	Expend physical energy for long periods
		Bodily activity	3. Agility	React quickly: has dexterity, coordination
		Sensory inputs	4. Vision	See details and color of objects
			5. Hearing	Recognize sound, tone, and pitch
	Mental	Vigilance and attention	6. Perception	Observe and differentiate details
			7. Concentration	Attend to details amid distractions
			8. Memory	Retain and recall ideas
		Information processing	9. Comprehension	Understand spoken and written ideas
			10. Problem-solving	Reason and analyze abstract information
			11. Creativity	Produce new ideas and products
	Learned	Quantitative computation	12. Numerical computation	Solve arithmetic and numerical problems
		Communications	13. Oral expression	Speak clearly and effectively
			14. Written expression	Write clearly and effectively
		Action selection and projection	15. Planning	Project a course of action
			16. Decision-making	Choose a course of action
		Application of information and skill	17. Craft knowledge	Apply specialized information
			18. Craft Skill	Perform a complex set of activities
	Motivational	Unprogrammed	19. Adaptability—change	Adjust to interruptions and changes
		Cycled	20. Adaptability—repetition	Adjust to repetitive activities
		Stressful Working	21. Adaptability—pressure	Adjust to critical and demanding work
		Secluded Conditions	22. Adaptability—isolation	Work alone or with little personal contact
		Unpleasant	23. Adaptability—discomfort	Work in hot, cold, noisy work places
		Dangerous	24. Adaptability—hazards	Work in dangerous situations
		Absence of direct supervision	25. Control—dependability	Work with minimum of supervision
		Presence of difficulties	26. Control—perseverance	Stick to a task until completed
		Unstructured conditions	27. Control—initiative	Act on own, take charge when needed
		Access to valuables	28. Control—integrity	Observe regular ethical and moral codes
		Limited mobility	29. Control—aspirations	Limit desire for promotion
	Social	Interpersonal contact	30. Personal appearance	Meet appropriate standards of dress
			31. Tolerance	Deal with people in tense situations
			32. Influence	Get people to cooperate
			33. Cooperation	Work as a member of the team

From "An Empirical Test of a Trait-oriented Job Analysis Technique" by F. Lopez, G. Kesselman, and F. Lopez, 1981, *Personnel Psychology, 34*, p. 484. Copyright 1981 by Personnel Psychology, Inc. Reprinted by permission.

whether these values were computed before or after raters were allowed to discuss and resolve differences in their trait ratings). Lopez et al. (1981) acknowledged that "some traits may be more reliably assessed than others" (p. 486); however, they did not report reliabilities for individual traits, and the seriousness of this problem for the TTA rating method has yet to be determined.

Other approaches to inferring KS and AO requirements do not rely at all on a prior analysis of job tasks or behaviors. For example, the adjective checklist approach method advocated by Arneson and Peterson (1986) requires job incumbents to simply "identify the kind of personal characteristics that make for an excellent worker" (p. 3). This is accomplished by checking personal trait adjectives (see Figure 13) that the incumbent feels are required; these ratings are then converted to personality trait specifications (i.e., intellectence, adjustment, prudence, ambition, sociability, and likability).

At a minimum, methods that directly rate KS and AO requirements must be able to demonstrate substantial agreement across raters regarding the traits deemed necessary. Unfortunately, the data reported by Arneson and Peterson indicated less-than-perfect agreement among their raters. One-hundred percent agreement on ratings of the trait-based adjectives (i.e., required vs. not required) was obtained on only 25 of the 80 items; using a less stringent 90 percent criterion of interrater agreement, only 59 of the 80 adjectives were agreed upon.

Even methods for identifying KS/AO specifications that *are* based on a rigorous job analysis may fail to produce acceptable interrater agreement. For example, Hughes and Prien (1989) first performed a detailed task analysis, which was followed by rating the KS/AO items listed in Table 5.

Although Hughes and Prien explicitly attempted to link the job analysis data to KS/AO ratings, their results indicated a distressing lack of agreement among judges. As the interrater agreement statistics reported in Table 5 indicate, there was considerable variability in trait ratings, and the overall level of agreement was quite low. Similarly disconcerting results were found when interrater agreement correlations were examined: Regarding the Importance ratings, the mean interrater r was only .31 ($S = .12$), and rs ranged from a low of −.03 to a high of only .50. Interater reliability judgments regarding the "difficulty to acquire" each KS/AO construct were even lower: The mean r was only .16 ($S = .34$), and values ranged from a low of −.40 to a high of .64. When one considers that these interrater agreement correlations are only sensitive to the *shape* of the profile of KS/AO requirement ratings, and *not* sensitive to differences across raters in terms of the *level* or *amount* of each trait needed (including that source of variance would likely further reduce the magnitude of interrater agreement), one can easily agree with Hughes and Prien that "the problem indicated by the interrater correlations is quite serious" (p. 287).

Ironically, Hughes and Prien (1989) scrupulously followed the prescriptions outlined in the *Guidelines:* (a) Raters were given a detailed job analysis (343 task statements); (b) job-knowledgeable raters were used as SMEs; and (c) a systematic step-by-step process of linking KS/AO employee specifications to the job analysis was attempted. If SMEs demonstrate massive interrater disagreement under these very favorable conditions, what can be expected when on-site job analysis data are not collected, or when raters judge even more abstract traits (e.g., personality or biodata items)?

Even if KS/AO rating methods that produce higher levels of interrater agreement can be developed, the truly important question—whether the inference that the trait is required for successful job performance is *valid*—cannot be addressed by interrater agreement statistics. In addition, such methods appear to implicitly assume that there is a *single* profile of

FIGURE 13

Adjective Checklist Approach to Rating AO Constructs

Hogan Descriptive Adjective Inventory

Directions: Think of the kinds of personal qualities that make for an excellent worker in this position. Listed below are a number of adjectives. Read each one and decide whether or not it describes this ideal worker.

		Yes	*No*			*Yes*	*No*
1.	Simple	☐	☐	28.	Assertive	☐	☐
2.	Calm	☐	☐	29.	Charming	☐	☐
3.	Careless	☐	☐	30.	Impatient	☐	☐
4.	Ambitious	☐	☐	31.	Good with numbers	☐	☐
5.	Talkative	☐	☐	32.	Pessimistic	☐	☐
6.	Irritable	☐	☐	33.	Inconsistent	☐	☐
7.	Well-educated	☐	☐	34.	Complacent	☐	☐
8.	Self-doubting	☐	☐	35.	Likes to be alone	☐	☐
9.	Responsible	☐	☐	36.	Friendly	☐	☐
10.	Unassertive	☐	☐	37.	Reads a lot	☐	☐
11.	Introverted	☐	☐	38.	Happy	☐	☐
12.	Tactful	☐	☐	39.	Dependable	☐	☐
13.	Slow learner	☐	☐	40.	Hard working	☐	☐
14.	Cheerful	☐	☐	41.	Show-off	☐	☐
15.	Unconventional	☐	☐	42.	Cold	☐	☐
16.	Forceful	☐	☐	43.	Complex	☐	☐
17.	Sociable	☐	☐	44.	Anxious	☐	☐
18.	Hot-tempered	☐	☐	45.	Careful	☐	☐
19.	Many interests	☐	☐	46.	Lazy	☐	☐
20.	Tense	☐	☐	47.	Quiet	☐	☐
21.	Team player	☐	☐	48.	Good natured	☐	☐
22.	Noncompetitive	☐	☐	49.	Poorly educated	☐	☐
23.	Reserved	☐	☐	50.	Confident	☐	☐
24.	Diplomatic	☐	☐	51.	Unreliable	☐	☐
25.	Poor memory	☐	☐	52.	Leaderlike	☐	☐
26.	Depressed	☐	☐	53.	Extraverted	☐	☐
27.	Impulsive	☐	☐	54.	Blunt	☐	☐

From "Using the 'Big Five' Personality Dimensions in Job Analysis" by J. Hogan and S. Arneson, April 1987. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA. Copyright 1987 by Robert and Joyce Hogan. Reprinted by permission.

TABLE 5

KS/AO Constructs and Reliabilities

Cohen's Kappa Values for Task to Job Skill Linkage Analysis

	Job Skill	*Kappa*	*Significance*
2.	Ability to listen	.2337	.0084
3.	Ability to read	.1988	.1074
8.	Mechanical comprehension	.4077	.0000
12.	Agility	.3122	.0026
13.	Leg strength	.4145	.0003
14.	Arm strength	.4498	.0000
15.	Upper body strength	.4609	.0000
16.	Ability to think logically	.1774	.0359
19.	Ability to visualize spatially	.3236	.0271
21.	Ability to remember	.0171	.8013
22.	Ability to follow oral orders	.0208	.7544
27.	Color vision	.0203	.8442
30.	Lower body stregth	.4615	.0000
31.	Total body stregth	.4720	.0000
40.	Ability to balance	.3407	.0158
42.	Arithmetic ability	.3759	.0750
43.	Ability to read gauges	.3887	.0171
44.	Hand dexterity	.1900	.0882
45.	Eye/hand coordination	.2080	.0569
48.	Ability to bend	.3580	.0001
49.	Ability to work in confined spaces	.2411	.1172
53.	Sense of direction	.3104	.0587
54.	Observation skill	.1403	.1252
57.	Knowledge of physics	.3679	.0195
58.	Endurance	.3649	.0012
60.	Ability to reach compartments	.4536	.0013
63.	Ability to apply facts	.0145	.8306
71.	Deduction	.2111	.0177
78.	Reaction time	.3836	.0044
79.	Ability to analyze information	.3337	.0066
81.	Ability to make decisions	.2737	.0026
85.	Knowledge of chemistry	.4461	.0143
86.	Hand strength	.2577	.0004
PI1.	Claustrophobia	.4098	.1529

From "Evaluation of Task and Job Skill Linkage Judgments Used to Develop Test Specifications" by C. Hughes and E. Prien, 1989, *Personnel Psychology*, 42, p. 289. Copyright 1989 by *Personnel Psychology*, Inc. Reprinted by permission.

KS and AO traits that will be required to achieve successful criterion performance; the possibility that a compensatory model in which low levels of some traits can be offset by high levels on other traits does not appear to be considered.

Assessing KSs and AOs in Individuals. The second aspect of using job analysis data in the course of dealing with KS/AO requirements involves assessing the degree to which individual job candidates or employees possess the constructs (i.e., an individual assessment issue), and/or developing assessment methods to quantify applicant standing on the constructs of interest (e.g., assessment centers or work samples). The role of job analysis data in this process depends on whether KS versus AO characteristics are assessed; in the case of KSs, job analysis data can be directly used when developing assessment techniques to measure these characteristics, such as in specifying the behaviors to be sampled in work samples, performance tests, or job knowledge tests. In the case of AO requirements (e.g., honesty, leadership, dominance, intelligence, and need for achievement), job behaviors are typically not directly used in the assessment process and standardized tests or assessment exercises usually do not employ job-specific content.

Identifying Job Similarity for Validation Purposes. Court decisions and the *Uniform Guidelines* indicate that a behavior-based job analysis should be used when assessing job similarity for consortium-validation and validity generalization purposes (i.e., demonstrating that the jobs in question perform essentially the same profile of activities under similar conditions). An alternative method to solve the problem of validating tests in small organizations was advocated by Guion (1965) and consists of grouping jobs based on their similarity with respect to *individual dimensions* of work instead of the full profile of job activities.

This *synthetic validity* strategy consists of (a) collecting job analysis data on several different jobs, (b) identifying work dimensions that are to some degree common across the jobs, and (c) validating tests for groups of jobs that share individual work dimensions—previous validation data may be used, or the organization can collect its own. In terms of the kind of job analysis data that should be used, the method must be able to identify common dimensions of work behavior; although a task-based analysis to identify common duty areas can be conducted, standard worker-oriented instruments offer an efficient means of identifying common worker-oriented items or job dimensions.

Compensation

Although there is no lack of litigation regarding the area of employee compensation (particularly based on the Equal Pay Act), in comparison to employee selection there are more degrees of freedom in using job analysis data to make compensation administration decisions. These uses take two main forms: (a) determining whether the Equal Pay Act criteria of equal skill, effort, responsibility, and working conditions are present when comparing jobs (i.e., a job classification question) and (b) using job analysis data to predict compensation rates.

Measuring Skill, Effort, Responsibility, and Working Conditions. In interpreting the Equal Pay Act, the courts are called on to decide whether jobs that currently receive differential pay require equivalent skill, effort, responsibility, and are performed under similar working conditions. If so, a violation of the act exists. Typically, the jobs in question differ in terms of their relative proportion of male and female incumbents.

Job analysis data have been called on to facilitate the determination of whether jobs require equal skill, effort, responsibility, and working conditions (e.g., see Henderson, 1979).

Unfortunately, the question of measuring abstract constructs like effort and responsibility has parallels to the problems faced in determining KS/AO requirements. That is, considerable abstraction is required to make the inferential leap from job behaviors to a listing of skill, effort, and responsibility requirements. As was the case for identifying KS and AO requirements, a popular solution is to have SMEs directly judge these traits; alternatively, the kinds of job dimensions measured by worker-oriented job analysis questionnaires can be categorized into skill, effort, responsibility, and working condition clusters to allow quantitative comparisons between jobs. This latter approach is preferable from the standpoint of being able to demonstrate job relatedness.

Measuring Compensable Factors. Standardized worker-oriented job analysis questionnaires have been highly successful in the area of compensation; by using the job dimension scores produced by these instruments as *compensable factors*, or the abstract job characteristics on which organizations base compensation decisions, impressive levels of prediction of market compensation rates have been achieved. Indeed, these instruments have achieved levels of predictive efficiency that are virtually unparalleled in other areas of industrial and organizational psychology.

Much of this research has involved the PAQ (e.g., McCormick, Mecham, & Jeanneret, 1977; Robinson, Wahlstrom, & Mecham, 1974), and more recently, the JEI (Harvey et al., 1988). Procedurally, jobs are analyzed using worker-oriented methods, general job dimension scores are produced, and these work dimension scores are used as predictors of market compensation rates for the jobs in question. Once market pay policies with respect to each job dimension are captured using multiple regression, a predicted compensation value termed *points* can be determined for any new job by simply applying the regression weights in new samples of jobs.

One of the most notable aspects of this research concerns the magnitude of the multiple correlations obtained using these policy-capturing techniques on compensation data: Using the relatively primitive JEI, *R*s in the .70s and .80s predicting existing compensation rates have been obtained (e.g., Harvey et al., 1988). Studies using the PAQ have reported even higher levels of predictive efficiency. For example, McCormick et al. (1977) reported an *R* of .85.

There have been a few clouds on this otherwise sunny horizon, however, particularly concerning the ability of worker-oriented instruments like the PAQ and JEI to predict pay rates for managerial, professional, and executive jobs. The PAQ's authors have agreed (e.g., McCormick et al., 1977; J. Mitchell & McCormick, 1979) that it lacks sufficient coverage of the work activities of higher-level jobs and does not predict pay rates for these jobs as effectively as it does for blue collar jobs. Other instruments (e.g., EXCEL, PMPQ, PDQ) have subsequently been developed to remedy the lack of coverage of managerial and executive work dimensions seen in general-purpose instruments like the PAQ and JEI.

Some have criticized the use of policy-capturing approaches in general, charging that they serve to capture and perpetuate sex-based discriminatory pay policies that may exist in the labor market (e.g., Schwab, 1984, p. 87). In defense of market-capturing techniques, however, one must note that instruments like the PAQ and JEI did not *create* the market conditions being captured by these regression coefficients; there is little reason to kill the messenger that simply describes prevailing market policies.

Of perhaps greater importance, when one considers the fact that most holistic rating methods used for compensation purposes (e.g., the Hay system, described by Bellak, 1984) rate jobs on highly abstract, nonbehavioral compensable factors (e.g., "know

how," "accountability"), the behaviorally verifiable job analysis item ratings that worker-oriented instruments use to define compensable factors represent a quantum improvement in measurement precision and defensibility. One can easily hypothesize that far more pay inequity is caused by reliance on sex-stereotype-influenced, non-job-related holistic compensable factor ratings than is caused by capturing market pay policies.

Finally, research (e.g., Harvey, 1985b) comparing unit weights and rational weights—which presumably cannot perpetuate past sex-based pay discrimination—with market-based policy-capturing equations has suggested that the jobs that benefit most from *not* using market policies tend to be stereotypically male jobs, such as heavy equipment operator or police officer. Stereotypically female titles—secretary, receptionist—often receive *fewer* points using nonmarket equations. Additional research is needed to determine the generalizability of these findings.

Summary

Many personnel applications (e.g., performance appraisal, training, work samples, performance tests, job descriptions, developing KS-based job specifications) require moderate- or high-specificity job data (tasks, behaviorally defined worker-oriented items, critical incidents), both to facilitate the development of quality products as well as demonstrate job relatedness. Other uses (e.g., compensable-factor compensation systems, test validation using more abstract AO-based constructs) that require a relatively task-insensitive common metric on which to compare jobs are better served by using low-specificity data (e.g., job dimension scores).

In view of the poor levels of interrater reliability—to say nothing of the ultimate issue of validity—that to date have characterized methods that infer KS/AO-based employee specifications on the basis of direct SME ratings, there is little reason for optimism regarding the future of such methods. Researchers should focus on identifying empirically defined links between the domain of job activity constructs and the domain of human KS and AO characteristics, using job component or synthetic validity methods, for example.

The Dimensionality of Work

Background

The central concept in worker-oriented job analysis is that a number of general dimensions of work behavior underlie all jobs and that even task-heterogeneous jobs can be described and meaningfully compared in terms of their scores on these dimensions. The majority of research directed toward identifying general dimensions of work behavior has been conducted by factor analyzing structured job analysis questionnaires composed of worker-oriented items (e.g., Cunningham et al., 1983; Harvey et al., 1988; McCormick et al., 1972). However, the identification of general work dimensions is an issue facing more behaviorally specific, task-oriented job analysis methods as well; for example, Fine's FJA seeks to compare even highly dissimilar tasks in terms of their involvement with a small number of very general work behaviors. In short, the issue of defining the dimensionality of work centers on the question of identifying *general job behavior constructs.*

Identifying the dimensionality of work behavior is critical for two reasons. First, describing jobs in terms of general work dimensions is important for personnel functions in which it is critical to identify behavioral similarities between jobs that may be masked by task technology differences. Second, efforts designed to identify the basic dimensions on which *people* vary, or to identify general "types" of people, have been carried out by psychologists concerned with identifying and measuring individual differences. The general dimensions of concern to such researchers include cognitive abilities (e.g., Guilford, 1967), physical abilities (e.g., Fleishman & Quaintance, 1984), personality traits (e.g., Hogan, 1990), and biodata (e.g., Owens, 1971).

Industrial and organizational psychologists have for many years been calling for research to link the dimensional system defining general work behaviors with the dimensional system defining general human characteristics or types. Dunnette (1976) characterized these two domains as "the two worlds of human behavioral taxonomies" (p. 477) and highlighted the theoretical and practical importance of being able to systematically match the characteristics of people with the characteristics of jobs.

Unfortunately, research on the human dimensionality side of the equation appears to have advanced much further than research aimed at defining the dimensionality of work behavior. Since the turn of the century, individual differences psychologists have been attempting to span the dimensionality of human traits, particularly cognitive and perceptual abilities (e.g., Spearman, 1927). Systematic empirical work dimensionality research did not begin to emerge until the 1960s (e.g., Cunningham, 1964; Jeanneret, 1969; Palmer & McCormick, 1961).

Task-oriented Work Dimensions

Functional Job Analysis. Although typically viewed as a task-oriented method, FJA has at its core a theoretical statement regarding the general dimensionality of work activities. FJA is based on the position that all tasks can be described and compared in terms of three fundamental work dimensions termed *worker functions:* Data, People, and Things. Figure 14 lists these dimensions and the hierarchy of activities hypothesized by Fine to define each.

By determining a task's involvement, termed its *functional level*, with respect to the Data, People, and Things dimensions, a common content and rating scale metric is obtained that allows even highly dissimilar tasks to be compared meaningfully. Although the way in which functional level ratings are computed in FJA is quite different than that used in worker-oriented instruments (i.e., FJA rates tasks and then summarizes across tasks to form overall Data, People, and Things scores, whereas worker-oriented instruments apply factor scoring coefficients to worker-oriented item ratings to obtain dimension scores), the conceptual foundation of FJA directly addresses the issue of identifying general dimensions of work activity. At a conceptual level, then, a job's overall functional-level ratings would comprise Type 3 data in the taxonomy of methods presented in Table 1.

At a practical level, however, because FJA's Data–People–Things dimensions were developed through largely nonempirical means, little is known about the measurement properties or construct validity of these dimensions. One study that addressed these issues (McCulloch & Francis, 1989) raised questions about the validity of the hierarchical ordering of activities defining these scales. Using confirmatory factor analysis of a pool of items measuring each of the seven levels of the People scale (e.g., persuading, supervising, mentoring), McCulloch and Francis found that although the seven constructs themselves were empirically justified, there was no support for the predicted hierarchical relationship among them.

Dimensionality of Task Inventories. A similar lack of information exists regarding the dimensionality of task inventory data. Given that task inventories typically include hundreds of individual task statements—making a factor analysis difficult using even

Strelau on Temperament and Personality

Strelau (1987) all but concedes the territory of personality to nontemperamentalists. He rightly notes that most personality theories—psychoanalytic, social learning, phenomenological, cognitive—have no use for temperament at all, with the notable exception of trait theories, which he explicitly excludes from his conceptual analysis. Noting that temperament has been regarded as one of the elements of personality, as a synonym of personality, and as a phenomenon not belonging to the structure of personality, he opts for the latter view. He discusses five respects in which there is at least a polar difference between temperament and personality:

1. Temperament is biologically determined, whereas personality is a product of the social environment (cf. Leontev, 1978).
2. Temperamental features may be identified from early childhood, whereas personality is shaped in later periods of development.
3. Individual differences in temperamental traits like anxiety, extraversion–introversion, and stimulation-seeking are also observed in animals, whereas personality is the prerogative of humans.
4. Temperament stands for stylistic aspects, personality for content aspect of behavior.
5. Unlike temperament, personality refers to the integrative function of human behavior. (What Strelau seems to say at this point is that personality is a teleological, temperament a causal concept).

The opposition between Strelau's view and the personological stance set out above is a great deal resolved by considering it as a semantic difference: Strelau himself acknowledges the close correspondence between the temperament and the personological, or trait-theoretical, or individual differences approach of personality for which I have argued in conformity with, most notably, Eysenck (Eysenck & Eysenck, 1985). There is a difference in strategy, however. To put it dramatically, personologists are busy trying to liberate what they consider to be their territory from alien invaders, whereas the temperamentalist, for fear of being himself overrun by the forces of the grand personality theories, has resigned himself to abandoning the territory of personality and retreating to his own stronghold, awaiting better times. My argument, in this admittedly dubious metaphor, is that the time has come to join forces with the resistance in the land of personality.

Let us examine Strelau's arguments against identifying personality with temperament. The first is that many personality psychologists would disagree with statements like Eysenck's (1986) that genetic factors are the main determinants of personality. There are two sides to this issue. One is the debate on heritability coefficients—their database, and their susceptibility to restriction of range: for example, the more the general level of measured intelligence goes up, reflecting the influence of environmental optimalization, the higher the heritability coefficient, because the contribution of environmental factors to individual differences has decreased correspondingly. Thus there may be substantive disagreement without rejecting the viewpoint as such. The other side is that reduction to biological factors ought not play a central role either in personology or in the study of temperament, as I shall argue more extensively below; consequently,

that the objection is also peripheral. Strelau's second argument is that personality is at least in part socially determined. I have spent the major part of this chapter arguing against that view. His third point is that a temperamental conception of personality can lead to socially harmful consequences such as racism. This is a tenuous argument, for two reasons: neglecting or denying temperamental differences can also be socially harmful; and, generally speaking, the misuse of scientific points of view is not prevented by abandoning them, since they will be revived by others if the need arises.

A secondary question, if the personological viewpoint is accepted, is whether all traits outside the intelligence domain are temperamental (as Buss & Plomin, 1984; Eysenck & Eysenck, 1985, state or imply) or whether temperament is a proper subset of personality. There are several ways to approach this issue. I shall use lexical analysis, which thus far has not been applied to this issue.

To select personality-descriptive terms from the lexicon (for a historical review of this research paradigm, see John, Angleitner, & Ostendorf, 1988) is to construct an implicit or extensional definition of personality. Such listings of trait terms are far more concrete than abstract, intensional definitions of personality. Ideally, we should have a listing of temperamental traits at our disposal, in addition to the lists of personality traits that have been made available (Angleitner, Ostendorf & John, 1989; Brokken, 1979; Goldberg, 1980). As no such list of temperament traits has yet been produced, only tentative and indirect comparisons can be made. I shall argue that temperament is the core of personality, that is: On the one hand, temperament does not coincide with personality, but is a proper subset of it; on the other hand, temperamental traits are a central subset rather than a peripheral one.

Five personality factors have emerged from the lexical research tradition: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect or Openness (Angleitner & Ostendorf, in press; Goldberg, 1980; McCrae & Costa, 1985). Temperamentalists (e.g., Buss & Plomin, 1984; Eysenck & Eysenck, 1985; Fey, 1978; Strelau, 1983; Zuckerman, 1979) have usually distinguished fewer dimensions, which can be conceived as rotations of two or more of these factors, or have made finer distinctions within dimensions. Especially the dimension of Intellect or Openness does not seem to belong to the temperament domain as defined thus far, even though there is some overlap between the temperamental variable of Sensation-Seeking and certain facets of Openness.

Interestingly, some temperamental variables emphasize the opposite pole of personality factors. Conscientiousness, for example, does not sound very temperamental, whereas its approximate polar opposite, Impulsivity, does; the same relation holds for Agreeableness and Aggressiveness. Such differences of emphasis do not impress one as very deep, but this conceptual issue may deserve further study.

With regard to the central position of temperamental traits, a finding by Brokken (1979) in his study of trait-descriptive adjectives is relevant. Brokken had subjects rate the 1203 Dutch adjectives on fundamentality, that is, the extent to which these terms were judged fundamental versus superficial in describing people's personalities. Over adjectives, he correlated this rating with the terms' scores on the two criteria on which they were selected. These criteria consisted of whether a term could be substituted in the following two framing sentences: (1) "He or she is a [adjective] person" (Person criterion), and (2) "He or she is [adjective] by nature" (Nature criterion). The fundamentality

Exhibit C

Whitehead, W. C. (1985). Clinical decision making on the basis of Rorschach, MMPI, and automated MMPI report data (Doctoral dissertation, University of Texas at Southwestern Medical Center at Dallas, 1985). *Dissertation Abstracts International, 46-08B*, 2828.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Hunsley et al, 2003 in
Lilienfeld, Lynn, & Lohr
Science and Pseudoscience in
Clinical Psychology, 2003

3

Controversial and Questionable Assessment Techniques

JOHN HUNSLEY
CATHERINE M. LEE
JAMES M. WOOD

At the heart of the scientific enterprise, including a scientific approach to psychological assessment, lie the principles of falsifiability and methodological skepticism (e.g., Alcock, 1991; Bunge, 1991; Popper, 1959). At a minimum, these principles require that claims about the scientific merits or validity of a hypothesis, measure, or theory be framed in a such a way that they can be (1) subjected to empirical investigation (i.e., by data-based investigation, rather than by reliance on appeals to anecdotal evidence or to special knowledge or authority), (2) refuted or disconfirmed by empirical investigation, and (3) independently investigated (and ideally replicated) by both proponents and critics of the claims. Moreover, there is a presumption in the scientific enterprise that the burden of proof to evaluate or demonstrate the validity of such claims rests with those who are making them (Lett, 1990). It is therefore incumbent on proponents of an assessment strategy to demonstrate empirical evidence of its reliability and validity and to supply norms for relevant populations.

In this chapter, we focus on a particular class of assessment approaches and techniques that are widely used, but considered questionable by most psychologists who adhere to scientific principles. The assessment techniques that we discuss may have some limited merit as indicators of psychological phenomena, but they are commonly used in clinical practice in a manner that goes well beyond what is appropriate or justifiable based on scientific evidence.

Exhibit D

PSYCHOLOGICAL TESTS

Psychological assessment is not synonymous with psychological testing (cf. Matarazzo, 1986), but psychological tests commonly comprise a major part of the assessment process. We therefore focus on specific problematic psychological tests. There are thousands of psychological tests, which vary enormously in their complexity and scientific merit. By definition, a psychological test is the measurement of a sample of behavior obtained under standardized conditions and that has established rules for scoring or interpreting this sample (Anastasi, 1988).

Standards for psychological tests and for their appropriate professional use are well-developed and widely known (*Standards for Educational and Psychological Testing*, 1985, 1999). These standards set out the criteria against which psychological tests are evaluated. They also serve to ensure that test developers and test users meet consensually defined expectations, held by the profession and the public, with regard to the scientifically appropriate use of tests. Proponents of questionable and controversial tests frequently claim the legitimacy that is associated with scientifically sound measures. However, they also sometimes deny that their test should be subjected to the high standards expected of a psychological test because it is not "really" a test, but merely a method for collecting information.

TEST CONSTRUCTION AND PSYCHOMETRIC PRINCIPLES

We next focus on elements that are required for a test to be both psychometrically sound and clinically useful. These elements, which hold for all types of psychological tests, are standardization (of stimuli, administration, and scoring), reliability, validity, and norms.

Standardization is essential for a psychological test, as it is the first step in ensuring that obtained results can be replicated by another assessor. Unless there is standardization, any results are likely to be highly specific to the unique aspects of the testing situation. Standardization is necessary to ensure that the influence of unique aspects of the testing situation and the assessor are minimized. To this end, test developers typically provide detailed instructions regarding the nature of the stimuli, administrative procedures, time limits (if relevant), and the types of verbal probes and responses to the examinee's questions that are permitted. Instructions must also be provided for the scoring of the test. In some cases, only simple addition of responses is required to obtain a test result; however, even in these cases, there is clear evidence that computational errors that compromise the validity of the results often arise (e.g., Allard, Butler, Faust, & Shea, 1995). For many tests, complex scoring rules may require that assessors receive extensive training to achieve proficiency in accurate scoring.

Reliability is the next criterion that must be addressed in the development of a scientifically sound test. The basic issue addressed by reliability is one of consistency—whether (1) all aspects of the test contribute in a meaningful way to the data obtained (internal consistency), (2) similar results would be obtained if the test was conducted and/or scored by another evaluator (interrater reliability), and (3) similar results would be obtained if the person was retested after the initial test (retest reliability or test stability). That is, standardization of stimuli, administration, and scoring are necessary, but not sufficient, to establish reliability. Reliable results are crucial for generalizing the results and their psychological implications beyond the immediate context of the assessment. Even thorough and complete test standardization cannot guarantee reliability. For example, the test may consist of too many components that are influenced by ephemeral characteristics of the examinee or by contextual characteristics of the testing, including demand characteristics associated with the purpose of the testing or the behavior of the examiner. Alternatively, the scoring criteria for the test may be too complicated or insufficiently detailed to ensure reliable scoring across different assessors.

Validity addresses the issue of whether the test measures what it purports to measure. A standardized and reliable test does not necessarily yield valid data. Validity is a matter of ensuring that the test samples the type of behavior that is relevant to the purpose of the test (content validity), provides data consistent with theoretical postulates associated with the phenomena being assessed (concurrent and predictive validity), and provides a measure of the phenomenon that is minimally contaminated by other psychological phenomena (discriminant validity). In applied contexts, an additional form of validity should be considered, namely incremental validity—the extent to which data from a test add to our knowledge over and above the information gleaned from other data (Sechrest, 1963). Although it is common to talk about a test as either valid or invalid, actual validity is far more complex. Many psychological tests consist of subscales designed to measure discrete aspects of a more global construct. In such situations, it is erroneous to talk about the validity of the test, because the validity of each subscale must be established. Moreover, global validity of a test or subscale does not exist because validity is always established within certain parameters, such that a test may be valid for specific purposes within specific groups of people (e.g., specific ages or genders). Finally, a test may be used for multiple purposes, but its validity for each purpose must be empirically established. For example, knowing that a self-report test of psychological distress is a valid indicator of diagnostic status does not automatically support its use for such forensic purposes as determining competency or child custody arrangements.

For a test to be clinically useful, it must meet the criteria of standardization, reliability, and validity. However, to meaningfully interpret the re-

sults obtained from a single individual, it is essential to have norms, specific criterion-related cutoff scores, or both (*Standards for Educational and Psychological Testing*, 1999). Without such reference points, it is impossible to determine the meaning of the test results. The results must be compared with some type of standard to have meaning: Knowing that a person scored low or high on a test (i.e., relative to the range of possible scores) provides no meaningful information. Comparisons must be made with either criteria that have been set for a test (e.g., a certain degree of accuracy as demonstrated in the test is necessary for the satisfactory performance of a job) or with population norms. Selecting the target population(s) for the establishment of norms and then actually developing the norms are challenging tasks. For example, are the norms to be used for comparing a specific score with those that might be obtained within the general population or within specific subgroups of this population (e.g., gender-specific norms), or are the norms to be used for establishing the likelihood of membership in specific categories (e.g., nondistressed versus psychologically disordered groups)? As with validity considerations, it may be necessary to develop multiple norms for a test, based on the group being assessed and the intent of the assessment.

DISTINGUISHING BETWEEN VALID AND INVALID USE OF ASSESSMENT TECHNIQUES: OR, WHEN IS A TEST NOT A TEST?

Building on these considerations, the appropriate use of psychometrically sound tests requires that guidelines for administration and scoring are followed and that relevant validity data and group norms are used to interpret the obtained data. Care must be taken to ensure that the test is valid for the assessment purpose at hand and for the person being assessed.

Widely accepted definitions of psychological tests should enable psychologists to determine whether an assessment technique should be considered as a test. Specifically, there are two necessary conditions for defining an information-gathering activity as a psychological test. First, a sample of behavior is collected in order to generate statements about a person, a person's experiences, or a person's psychological functioning. Second, a claim is made or implied that the accuracy or validity of these statements stem from the way in which the sample was collected (i.e., the nature of the stimuli, technique, or process that gave rise to the sample of behavior), not just from the expertise, authority, or special qualifications of the assessor. When both conditions are present, we consider that the process used to collect and interpret the behavioral sample is a psychological test. This is consistent with the recently revised standards for psychological tests, which emphasize that an assessment method that relies on or uses the concepts and techniques of psychological testing is a test (*Standards for Educational and Psychological Testing*, 1999). Despite the apparent clarity of this definition, proponents of questionable techniques sometimes employ the term "psychological test" loosely, arguing simultaneously that scientific and professional standards, expectations, and responsibilities are inapplicable to their techniques, while also claiming a valid approach that is supported by scientific evidence. Thus, the issue of whether a specific form of data collection constitutes a test is far from an ivory tower concern for semantic hair-splitting.

For example, the Rorschach Inkblot Test has been used by clinical psychologists for many decades. Some Rorschach proponents have argued that the Rorschach is not a psychological test at all, but that it is instead a method of interviewing that generates data relevant to the practice of clinical assessment (e.g., Aronow, Reznikoff, & Moreland, 1995; Weiner, 1994). A slight variation on this approach is to treat the Rorschach as a test (by using a recognized scoring system), but then "enriching" the test results with personalized, interpretive speculations stemming from selected aspects of the examinee's test responses (e.g., Acklin, 1995; Fischer, 1994). Such positions allow the Rorschach assessor to claim that there is scientific evidence supporting the use of the Rorschach, while simultaneously freeing the assessor to use the data in a manner unconstrained by issues of administration, norms, reliability, or validity. An example of the problem associated with failing to recognize a test *as* a test is contained in a recent guide to conducting child custody evaluations (Ackerman, 1995). The author recommended that a Rorschach protocol be scored with Exner's (1993) Comprehensive System (a scoring system which, as we will demonstrate in a subsequent section, has some limited scientific merits) to comply with professional standards. However, he then contradicted this position by suggesting that it was not always necessary to score a Rorschach protocol, for an experienced clinician can allegedly assess anxiety, depression, and thought disorder on the Rorschach without going through the rigor of formal scoring (p. 116). This type of contradictory reasoning is common among some users of controversial tests.

CONTROVERSIAL AND QUESTIONABLE ASSESSMENT TECHNIQUES: SOME EXAMPLES

At this point, we consider a number of problematic assessment techniques that, according to surveys of assessment practices, are used by large numbers of clinical psychologists. Of course, there are many other examples of unscientific psychological assessment techniques that one could consider (e.g., for a thorough critique of graphology, see Beyerstein & Beyerstein, 1992). However, we have chosen to focus on five questionable assessment

techniques that continue to be routinely used by psychologists: the Rorschach Inkblot Test, the Thematic Apperception Test (TAT), projective drawings, anatomically detailed dolls (ADDs), and the Myers–Briggs Type Indicator (MBTI).

Psychologists in the domain of personality assessment have long distinguished between projective tests and self-report inventories (Anastasi, 1988). Projective tests such as the Rorschach or TAT generally present the person being tested with an ambiguous stimulus (such as an inkblot or a picture without a caption) and ask for an open-ended response to the stimulus (e.g., "What might this be?" or "What do you think is happening in this picture?"). In contrast, self-report inventories generally present the examinee with a statement (e.g., "I often feel like crying") and ask the person to choose among two or more responses to indicate the extent to which the statement accurately reflects the person's experience.

Among problematic techniques discussed in this chapter, the majority fit the definition of projective techniques. As we will demonstrate, problems of standardization are rife among projective techniques. Questionable techniques are not, however, limited to projectives. To illustrate this point, we review a self-report personality measure, the MBTI. Although standardization is not an issue with this test, concerns about its reliability and validity highlight the need for clinicians to select tests that have firm scientific support.

THE RORSCHACH INKBLOT TEST

The Rorschach Inkblot Test consist of 10 cards, each containing symmetrical inkblots, some in color and some in black and white. Examinees are asked to report what they see in these ambiguous stimuli. According to Rorschach proponents, important evidence regarding psychological functioning can be obtained from the Rorschach when one considers the nature of what is seen, what aspects of the card are used in the responses, the sequence of responses given during testing, and even the examinee's nonverbal reactions to the inkblots. For much of the 20th century, several distinct approaches to the administration and scoring of the Rorschach existed, and many clinicians tended to use elements of different systems and to "personalize" the scoring and interpretation of the Rorschach based on their own experiences (Exner & Exner, 1972). However, Exner's Comprehensive System (CS; Exner, 1974, 1993) has become the prevailing approach to teaching and researching the Rorschach (Hilsenroth & Handler, 1995; Shontz & Green, 1992). Indeed, it has been suggested that the CS should now be considered the primary scoring system in evaluating the scientific status of the Rorschach (Weiner, 2001). Exner's approach focuses on structural elements of Rorschach responses (i.e., the specific features of the inkblot that are involved in the response) and emphasizes the need to use appropriate scientific data in interpreting the test. In contrast to many Rorschach proponents (e.g., Weiner, 1994), Exner has always insisted that the Rorschach is a psychological test and, as such, must meet the standards expected of a test (e.g., Exner, 1997).

As the most commonly researched and used projective measure (Butcher & Rouse, 1996; Watkins, Campbell, Nieberding, & Hallmark, 1995), the Rorschach has been the focus of a great deal of scientific attention over the past 50 years. Even Rorschach proponents have accepted the fact that most early research (i.e., before 1970) was so poor that it should not be regarded as offering evidence for the validity of the test (Exner, 1986). As a result of the ascendancy of the CS, this Rorschach scoring method has been the subject of increased scientific scrutiny from both proponents and critics of the Rorschach. It is to this evidence that we now turn.

Standardization

The Rorschach is a complex measure to administer, score, and interpret, requiring a modal time of 3 hours of clinician time (Ball, Archer, & Imhof, 1994). The CS offers very clear information on administration and scoring, with extensive tables and computer software to aid in the interpretation of the test results. Assessors are provided directions on the seating arrangement to be used, the sequence of card administration, the instructions to be given to examinees, and the permissible type of assessor probes and responses to questions. The CS also requires that responses to all cards be obtained before the assessor queries the examinee, response by response, on the elements of the card that influenced the responses of the examinee (known as "determinants").

As the evidence from surveys on the clinical use of the Rorschach suggest that the majority of clinical psychologists use the test, at least occasionally, it is surprising that there are no data on the extent of assessors' fidelity in following CS administration and scoring requirements. Because many graduate courses on the Rorschach include information on multiple scoring systems, and about one third of these courses do not even teach the CS (Hilsenroth & Handler, 1995), there is no reason to believe that the traditional tendency of Rorschach users to "borrow" scores from different scoring systems and to personalize the scoring has diminished. This is a key issue in evaluating the scientific basis of the Rorschach in clinical use, for without evidence that Rorschach data were obtained in a fashion consistent with CS requirements, research stemming from the CS cannot be used to support the assessor's interpretation of the examinee's responses. Even if CS administrative requirements are scrupulously followed, there is extensive evidence that relatively innocuous contextual factors in Rorschach ad-

Everson and Boat (1994) illustrated the memory stimulus function of ADDs: "A clear example . . . is the case of a young child who, after noticing the pattern on the male doll's underwear stated, 'Granddaddy's underpants had hearts on them.' " (p. 117). This is an excellent example of how clinicians who are sensitized to child sexual abuse could misconstrue an ambiguous response. It is possible that the child has seen grandfather in his underwear, without there being any sexual abuse, or that the child has seen grandfather's underwear without him wearing it.

The American Psychological Association (APA) has commissioned two task forces to examine the validity of ADDs. The first concluded that although the dolls are not standardized and although there are no normative data and no uniform standards for conducting interviews, doll-centered assessment "used as part of a psychological evaluation and interpreted by experienced and competent examiners may be the best available practical solution for a pressing and frequent clinical problem (i.e., the investigation of the possible presence of sexual abuse of a child)" (American Psychological Association, 1991, p. 722). Furthermore, the task force exhorted psychologists who undertake doll-centered assessment to be competent (although this was not defined), to document their procedures, and to provide clinical and empirical rationales for their procedures and interpretations. These recommendations reflect a puzzling mixture of reliance on unspecified clinical wisdom coupled with reference to a research literature that the task force concluded was nonexistent.

The Anatomical Doll Working Group, funded by the APA (Koocher et al., 1995), reiterated the conclusion of the first task force, noting that ADDs do not meet any of the criteria for a valid psychological test or a projective technique. Koocher and colleagues (1995) advised that conclusions about child sexual abuse cannot be made on the basis of doll play alone and that reports of children under 4 years of age are particularly prone to be affected by misleading questions. These cautions notwithstanding, Koocher and colleagues reasserted the original APA task force position. Both APA resolutions reflect the tensions in psychological practice and the lip service paid to science by some psychologists who are willing to examine research literature, but equally willing to dismiss it if it does not correspond to views founded on their clinical experience.

Various reviewers have concluded that it is not clear what children's sexualized play with ADDs indicates (Babiker & Herbert, 1998). Critics charge that the incremental validity of ADDs must be demonstrated (e.g., Ceci & Bruck, 1995; Wolfner et al., 1993). That is, they argue that ADDs must be shown to consistently *add* to our ability to determine whether a child has been abused above and beyond already available information, such as interviews, observations, and rating scales. Advocates argue that ADDs are no worse than other assessment strategies (Aldridge, 1998; Koocher et al., 1995). This latter view stands in sharp contrast to the

prototypical scientific position we noted earlier, namely that the onus of demonstration of the utility of a particular assessment strategy rests with its proponents.

Conclusion

Many proponents of ADDs adopt scientific language by referring to "evidence," "studies," "research," and "empirical support." However, they seek to absolve the procedures from the scrutiny of scientific standards by denying that ADDs constitute a psychological test. Paradoxically, some proponents promote their approaches as scientifically supported while rejecting arguments that these measures be held to scientific standards. We reiterate previous findings that neither the stimuli nor the procedures used in ADD assessments are standardized. Given this lack of standardization, it is impossible to collect normative data on the behavior of abused and nonabused children. We reject claims that ADDs can be used as screening instruments without meeting the standards for psychological tests and therefore strongly advise against their use for this purpose in investigations of child sexual abuse.

MYERS–BRIGGS TYPE INDICATOR

The Myers–Briggs Type Indicator (MBTI; Myers & McCaulley, 1985) is a self-report test based on Jung's personality theory. Jung's theory of personality types, designed to be a comprehensive account of personality functioning, posits four basic personality preferences that are operationalized in the MBTI as bipolar, continuous constructs: extraversion–intraversion (oriented outwardly or inwardly), sensing–intuition (reliance on sensorial information versus intuition), thinking–feeling (tendency to make judgments based on logical analysis or personal values), and judgment–perception (preference for using either thinking–feeling or sensing–intuition processes for interacting with the world). Based on scores obtained for these four dimensions, established cutoff scores are used to assign examinees to one of 16 different personality type categories (e.g., extraverted, sensing, thinking, judgment). The use of these 16 categories has been controversial, as they are consistent with neither Jungian theory nor data gathered from the MBTI (Barbuto, 1997; Garden, 1991; Girelli & Stake, 1993; Pittenger, 1993).

The MBTI is available in several versions, although the standard version is a forced-choice 126-item test. During the past two decades, the MBTI has been translated and normed in many languages, and it is among the most commonly used measures of normal personality (McCaulley, 1990). Although it was developed to be useful in education,

counseling/therapy, career guidance, and workplace team-building, it is within the assessment practices of career guidance and personnel selection that the MBTI has gained dominance, so much so that it is routinely used by psychologists for gathering information on possible career paths and job placement (Coe, 1992; Jackson, Parker, & Dipboye, 1996; McCaulley & Martin, 1995; Turcotte, 1994). The research on the MBTI is impressive in scope, with hundreds of studies published in the past two decades on a range of personality, educational, and vocational constructs.

Standardization

Given that the MBTI is a published self-report test, standardization of test instructions and test items should be assured when used appropriately. Several short form versions of the test are available. However, because of concerns about limited comparability vis-à-vis the full test (Harvey, Murry, & Markham, 1994), use of these short forms is not recommended.

The MBTI manual (Myers & McCaulley, 1985) provides detailed instructions on scoring the MBTI and converting the scores to 1 of the 16 personality types. Information is also provided on how the results of the test should be interpreted, both in general terms and with reference to counseling, educational, and career counseling contexts. One of the problems with the use of the 16 types that has consistently emerged in the research literature is the appropriateness of the cutoffs used to assign examinees to a type. Researchers have found that scale scores close to the cutoffs frequently lead to classification errors. To address this issue, there have been calls for changes in response format and scale scoring (Girelli & Stake, 1993; Harvey & Murry, 1994; Harvey & Thomas, 1996; Tzeng, Ware, & Chen, 1989).

Norms

The normative information reported in the manual is based on data from tens of thousands of research participants. Normative data for men and women are available across a wide range of ages (15–60+ years) and occupations. Research has tended to support the appropriateness of these norms and to find them applicable across minority groups and cultures (e.g., Kaufman, Kaufman, & McLean, 1993). However, as the data for minority groups are limited and there is some question about the interpretation of age-related factors influencing these data (Cummings, 1995), users of the MBTI would be well advised also to consider data from a recent study designed to obtain normative data from a representative sample of American adults (Hammer & Mitchell, 1996).

Reliability and Validity

The MBTI has typically been found to exhibit acceptable levels of internal consistency and test–retest reliability (Carlson, 1985). However, these data typically focus on the reliability of the four preference scores (e.g., extraversion–intraversion), and far less evidence is available on the reliability of the 16 types..

Dozens of validity studies of the MBTI are available in the scientific literature. They tend to focus on relating MBTI preference scores and types to a myriad of personality constructs, ability measures, and occupations. Nevertheless, there has been no real attempt to integrate the data from these studies to guide the valid interpretation of the test results. Moreover, there is relatively limited information on the predictive validity (i.e., whether accurate predictions of educational and career choices can be made on the basis of the MBTI) or the incremental validity and utility (i.e., whether the MBTI meaningfully adds to the prediction of these decisions; whether there are optimal educational, career, or employment decisions made on the basis of MBTI data) of the MBTI.

One aspect of the MBTI that has received extensive attention in the literature is the validity of the 16 personality types. Using a range of analytical procedures, including exploratory factor analysis, confirmatory factor analysis, and cluster analysis, researchers have generally found that (1) the observed factor structure of the MBTI is consistent with the hypothesized four personality preferences (Thompson & Borrello, 1986; Tischler, 1994), although often at a less than optimal level (Harvey, Murry, & Stamoulis, 1995; Jackson et al., 1996; Sipps, Alexander, & Friedt, 1985), and (2) the fit between the hypothesized 16 types and actual test data is poor (Lorr, 1991; Pittenger, 1993; but see also Pearman & Fleenor, 1996).

As a measure of global personality, the MBTI has been criticized for its failure to relate to other well-established vocational and personality measures. Although the efforts of the test developers to include concurrent validity data with a range of such measures in the test manual is laudable, there is little consistent evidence that the four personality preferences relate to comparable constructs assessed by other measures. Published research suggests that the MBTI bears little correspondence to measures of vocational preferences and job performance (e.g., Apostal & Marks, 1990; Furnham & Stringfield, 1993). Additionally, as a measure of global personality, the MBTI has been found to have limited correspondence with either of the two prevailing scientific models of personality structure, namely Eysenck's three-factor model and the five-factor model (Furnham, 1996; McCrae & Costa, 1989; Saggino & Kline, 1996; Zumbo & Taylor, 1993; but see MacDonald, Anderson, Tsagarakis, & Holland, 1994). One can

only conclude that the MBTI is insufficient as a contemporary measure of personality.

Conclusions

The MBTI is based on a explicit theory of personality, was developed and normed in a manner consistent with current standards, and has typically been found to be reliable at the level of the four personality preferences. However, questions about the reliability and validity of the 16 personality types and evidence of limited correspondence between the MBTI and other global measures of personality and vocational interests render the test suspect as an assessment tool. In the absence of a major revision of the test that addresses these shortcomings, psychologists are advised to rely on personality and vocational interest tests that have a sounder empirical basis (cf. Boyle, 1995).

CONCLUSIONS AND RECOMMENDATIONS

Psychologists face a daunting task in making sense of the vast literature on psychological assessment. In considering ways to address pressing clinical questions, they have access to a panoply of potential tools. Unfortunately, there are no simple ways to determine whether a test is scientifically valid. The fact that it is marketed in a prestigious professional newsletter or described in a scholarly journal provides no guarantee that it meets adequate test standards. For example, a recent meta-analysis (West, 1998) provided data that appeared to support the use of projective techniques in detecting sexual abuse in children. However, reexamination of the data revealed that the author had included in her calculations of effect sizes only significant findings, thereby inflating the apparent power of projectives to identify abused children (Garb, Wood, & Nezworski, 2000). Reference to research citations may therefore be inadequate in determining whether a test is appropriate. Many of the articles we reviewed contained abstracts claiming support for an assessment approach, although the article itself offered at best mixed support.

We urge psychologists to use as a reference the most recent edition of the *Standards for Educational and Psychological Testing* (1999), which restate and expand on the principles that have guided a scientifically based approach to assessment. Psychologists are required to engage in reasoned decision making as they select assessment tools. Knowledge of an assessment procedure may become obsolete as it is replaced by a more sophisticated understanding of its limitations. Psychologists using assessment procedures that lack a published manual are required to conduct their own scholarly review of the pertinent literature to determine whether the test meets basic standards of reliability and validity, and whether suitable norms exist. Those using published materials are also required to familiarize themselves with the data relevant to the use of the test. The fact that a published manual exists does not guarantee that the test meets standards of reliability and validity or that appropriate norms are available. Professional and ethical standards indicate that each psychologist is responsible for determining (1) the question that the psychological assessment is designed to answer, (2) whether there is a test that is adequately standardized and that yields reliable and valid information, and (3) whether pertinent norms allow interpretation of the responses on the test in a given circumstance. We caution psychologists against uncritically accepting the argument that a given assessment procedure is absolved from the obligation to meet accepted standards.

Among the tests we reviewed we found scant support for the Rorschach, some promising avenues for the TAT (although no support for this measure as it is currently used in clinical practice), only very limited promise for holistic scoring of some projective drawings, no support for ADDs as a screening instrument for evidence of sexual abuse, and evidence that the MBTI is a potentially reliable measure that lacks convincing validity data. A lack of standardization in the use of many of these techniques and an overreliance on unsubstantiated beliefs that certain people possess special interpretive powers (see also Chapter 2) has thwarted the possibility of advancing these techniques into the realm of scientifically supported assessment strategies.

Responsibility for demonstrating the adequacy of an assessment procedure rests first with those developing the procedure, and second with psychologists electing to use it. Proponents of specific assessment procedures are responsible for elucidating a standardized protocol that fully explains administration procedure. In the case of the Rorschach, TAT, projective drawings, and ADDs, it is necessary that proponents of each approach reach consensus on standardized administration and scoring. The maintenance of idiosyncratic versions of these tests mitigates against their establishment as scientifically sound assessment strategies.

Having established a standard protocol, research must address issues of reliability. Decision rules must be established that permit independent raters to reach the same judgments concerning a response. Once it has been established that a technique can be administered in a standardized fashion and the examinee's responses judged consistently, the issue of validity can be addressed. Tests that purport to measure a construct that cannot be independently measured by other tests or other assessment techniques are inherently unfalsifiable and therefore unscientific (see also Chapter 1). Finally, norms must be developed so that scores for an individual can be